



Australian Government
Bureau of Meteorology



Biodiversity Profiling

Components of a continental biodiversity information capability

A joint study between the Atlas of Living Australia, CSIRO Ecosystem Sciences
and the Bureau of Meteorology



VISIT...

LANZAROTE
Caliente.COM

Biodiversity Profiling: Components of a continental biodiversity information capability
Environmental Information Program Publication Series No. 2
ISBN: 978-0-642-70639-3

Environmental Information Services
Bureau of Meteorology
Tel: (02) 6232 3542
Email: environment@bom.gov.au
www.bom.gov.au

Citing this publication

Zerger, A., Williams, K.J., Nicholls, M., Belbin, L. Harwood, T., Bordas, V., Ferrier, S., and Perkins, G. 2013. *Biodiversity Profiling: Components of a continental biodiversity information capability*. Environmental Information Program Publication Series, No. 2, Bureau of Meteorology, Canberra, Australia. 72 pp.



With the exception of logos or where otherwise noted, this report is licensed under the Creative Commons Australia Attribution 3.0 Licence. The terms and conditions of the licence are at: <http://creativecommons.org/licenses/by/3.0/au/>

Cover: Charcoal Cups (*Anthracobia muelleri*), a fungi that commonly occurs after fire. © Alison Pouliot—used with permission. The Creative Commons licence does not apply to this material. If you wish to further use this material you must contact the copyright owner directly to obtain permission.

Contents

1. Executive summary.....	1
2. Introduction.....	3
2.1 Significance.....	3
2.2 Objectives	3
2.3 Assumptions and caveats	4
2.4 Organising framework and report outline	4
3. Enabling accessible data.....	6
3.1 Background: the Atlas of Living Australia	6
3.2 Data aggregation.....	7
3.3 Key principles	8
3.4 Including a resource in a data warehouse.....	10
3.5 Recommendations	29
4. Atlas of Living Australia (ALA) data and tools	31
4.1 Introduction.....	31
4.2 ALA users	31
4.3 ALA data types.....	31
4.4 Spatial coverage	32
4.5 Temporal coverage	35
4.6 Methods for data entry	35
4.7 Fitness-for-use / data quality	36
4.8 Sensitive data	37
4.9 Environmental data	38
5. Business requirements elicitation.....	39
5.1 Introduction.....	39
5.2 Workshop objectives.....	39
5.3 Detailed method	39
5.4 Discussion and recommendations.....	40
5.5 Methods recommendations.....	42
5.6 Thematic focus for biodiversity profiling	42
6. Case study – biodiversity profiling	43
6.1 Introduction.....	43
6.2 Modelling spatial pattern in biodiversity composition	46
6.3 Continental-scale measures of habitat condition	52
6.4 Estimating change in vegetation condition through remote sensing	56
6.5 Inference of biodiversity change	57
6.6 Conclusion	60
7. Conclusions.....	61
8. Acknowledgments	64
9. References.....	65

List of tables

Table 1. Study organising framework showing key components.....	5
Table 2. Records by States and Territories.....	35
Table 3. Temporal coverage of records and species for all species and Eucalyptus	36
Table 4. Example initial-scan business requirements showing against policy/program drivers.....	40
Table 5. Basic data profile attributes	48
Table 6. Key components of the vetting process	49
Table 7. Tabulation of results (see authors for details)	49
Table 8. Fitted GDM models	51
Table 9. Comparison of data to infer habitat change or vegetation condition at continental scales.....	55

List of figures

Figure 1. Atlas records by Kingdom.....	32
Figure 2. DNA records in the Atlas.....	32
Figure 3. Atlas records by life-form	32
Figure 4. Atlas records by species names	33
Figure 5. Atlas records by basis of records	33
Figure 6. Identification keys in the Atlas	33
Figure 7. Spatial coverage of Atlas records.....	34
Figure 8. Occurrence density in Australian region	34
Figure 9. Temporal coverage of records and species for all species	37
Figure 10. Five key priority elicitation phases.....	41
Figure 11. A hierarchy of options	44
Figure 12. Options relating to the 'how?' of biodiversity monitoring	44
Figure 13. A decision framework for evaluating biological records	47
Figure 14. Example maps for Kingdom Fungi.....	50
Figure 15. Example classification derived from GDM modelling.....	52
Figure 16. Derived woody-vegetation condition layers for 1975 and 2006.....	58
Figure 17. Change in retention of compositional diversity	59
Figure 18. Time series of aggregated (weighted average) p values for selected IBRA bioregions.....	60

1. Executive summary

Historically in Australia, some environmental domains have been relatively well served by operational environmental monitoring, and in some instances, forecasting capability. These include systems for weather, climate and water (Bureau of Meteorology), earth science information (by Geoscience Australia) and aspects of land cover monitored by synoptic remote sensing technology. However, in environmental domains such as biodiversity, it is recognised there is a general lack of consistent information available at continental scales to support decision-making. Through a collaboration with the Atlas of Living Australia (ALA), the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Bureau of Meteorology under the auspices of the National Plan for Environmental Information Initiative, this study explores key components for developing a national biodiversity monitoring capability. In particular, it looks at whether an approach that couples natural history collections with mapping data can be used to derive metrics of inferred change in biodiversity status across Australia, or profiles. The term profiling refers to our ability to describe biodiversity change through time and across space.

The Atlas of Living Australia's integration of the nation's natural history and related environmental data provides a unique opportunity to develop approaches for continental-scale biodiversity profiling. By combining this observational capability with modeling and mapped data such as that showing changes in vegetation extent acquired from satellite imagery, opportunities exist for developing continental scale metrics of biodiversity change through space and time. A key principle of this study was to use only existing operational capability in regard to data, modeling and contextual spatial information such as time series satellite imagery acquired to support the National Carbon Accounting System. In this manner, the approaches and findings remain operationally feasible at continental scales.

In addition to evaluating feasibility, the study pays

particular attention to the key 'soft' enablers for building such a national capability by reporting on the lessons learnt in establishing components of the ALA. This includes an exploration of federated data acquisition approaches, licensing, user-needs evaluation to stakeholder engagement and collaboration. Many of the lessons reported are generic and are likely to be useful for other environmental information activities that use a federated model for data acquisition.

An applied case study demonstrates that data aggregated by the ALA could be readily obtained and filtered to improve data quality as an input, along with spatial environmental data, to continental models of biodiversity. A condition time series was then derived from remotely sensed metrics of forest (woody vegetation) cover produced to support the National Carbon Accounting System. These outputs are limited to where forest cover is defined and may not distinguish fully between native and non-native woody vegetation, but overall resulted in a consistently plausible condition metric. The condition metric, combined with the models of biodiversity pattern, were used to assess the proportional loss or retention of biodiversity in the landscape. Over the time series of the condition data, patterns of change can be detected across individual regions, demonstrating the potential for linking data and models for continent-wide biodiversity monitoring.

The study observes that there is no one way biodiversity should be monitored in Australia and provides a conceptual framework to contextualise the approach adopted. Different policy, assessment and decision-making needs will place different demands on biodiversity monitoring, in addition to the fact that emerging technologies and the progress of science will enable the application of new approaches. Thus the case study is intended to illustrate the potential use of ALA data in monitoring, and therefore focuses on an approach particularly suited to such data. This approach could thus form one component of a broader integrated program.

The report ends with a treatment of generic conclusions and recommendations that have emerged from the establishment of the ALA and findings from the study herein given their relevance beyond only the biodiversity domain. The most notable of these include the following:

- The establishment of the Atlas of Living Australia and related-initiatives (e.g. TERN, IMOS, ANDS) and enduring Australian Government agencies and programs, highlights the importance of establishing a trusted central authority for coordinating the specification and implementation of standards to collect, integrate and disseminate information.
- The role of standards is critical for supporting both data aggregation and assimilation, in addition to dissemination and re-use.
- Aggregating data from multiple providers will demand flexible information architectures that accommodate both centralised and distributed approaches to data management and re-use.
- Metadata plays a critical role in supporting effective use and re-use of environmental information, particularly in the case of initiatives that operate as aggregators.
- Establishing a culture of continuous environmental data availability and free and open data access is likely to provide new opportunities to support more informed decision-making. This includes the development of common data licencing frameworks.
- There is a need to define standards for data capture and data storage to enable more efficient data acquisition at source, and to support interoperability and re-use, for example through definition of collection protocols (observing methods) and domain appropriate information models.

New opportunities are rapidly emerging for overcoming some of the earlier barriers to establishing a continental biodiversity profiling capability. The advent of national capabilities such

as the Atlas of Living Australia and the Terrestrial Ecosystems Research Network have rapidly improved our ability to monitor our landscape. New opportunities for data assimilation that integrate these capabilities with existing operational systems such as those maintained by government agencies including the Bureau of Meteorology, the Department of Sustainability, Environment, Water, Population and Communities and Geoscience Australia provide new opportunities to better understand ecological change to support continental-scale decision-making.

2. Introduction

Under the auspices of the National Plan for Environmental Information (NPEI) Initiative, the Bureau of Meteorology began targeted activities to explore the benefits of innovative approaches to environmental information acquisition, discovery, delivery and re-use, to provide new insights that address key Australian government policy needs. These activities build technical capability and explore the importance of non-technical enablers for environmental information acquisition, discovery and delivery. A number of these will be delivered through collaborations, particularly where mature capability already exists.

Some environmental domains are relatively well served by operational monitoring systems such as those for air and water (by the Bureau of Meteorology) and earth science information (by Geoscience Australia). However, in domains such as biodiversity and soils, there is a lack of consistent information available at continental scales to support environmental decision-making. The Atlas of Living Australia's (ALA) integration of the nation's natural history and related environmental data provides a unique opportunity to develop approaches for continental-scale biodiversity profiling. By combining this new capability with the ecological modelling methods developed by CSIRO Ecosystems Sciences, there is a unique opportunity to develop spatially-explicit measures of change in the state of biodiversity at continental scales to support government policy and decision-making. The term profiling in the context of this study refers to our ability to describe biodiversity change through time and across space. This project brings together the richness of the Atlas of Living Australia's biodiversity and environmental data holdings, the ecological modelling and synthesis expertise of CSIRO Ecosystems Sciences (components of which are implemented in the Atlas) and the detailed understanding of government policy needs and technical expertise in operational environmental information management of the Bureau of Meteorology.

2.1 Significance

This study is thematically important given the historical lack of consistent biodiversity monitoring capability in Australia. Relative to other physical assets such as water, atmosphere and oceans, our ability to monitor biodiversity at continental scales is limited. The advent of the ALA and its integration of a wide range of biological records and related environmental layers provide, for the first time, a unique national opportunity to progress this requirement. This study is also technically important as the maturity of the ALA provides important lessons for the NPEI Initiative on 'soft' enablers (for example, working with providers and licensing). Developing these enablers is fundamental to effective operational environmental information delivery. Further, the ALA is uniquely placed to guide future work given their reliance on the widest range of biological and related environmental information. This is unique among the suite of other similar Australian environmental monitoring capabilities.

From a 'business needs' perspective, the advent of the NPEI Initiative also provides a opportunity to bridge the 'divide' between research capabilities and solutions and Australian government decision-making requirements for environmental information. The initiative has recently completed a strategic activity to identify the requirements of Australian government that have a dependency on environmental information (http://www.bom.gov.au/environment/Statement_AGREI_web.pdf, last accessed 28 April 2013). One of the nine areas of interest identified through this process was biodiversity thus providing important context for this biodiversity profiling study.

2.2 Objectives

The objectives of the study were to:

- demonstrate the potential of using federated biological data (that is, the Atlas of Living Australia holdings), spatially explicit environmental

variables, and spatial modelling to derive metrics of inferred change in biodiversity status across the Australian continent;

- examine and report on the role of ‘enablers’ for developing an operational biodiversity monitoring system (for example, federated data acquisition, licensing, user-needs evaluation, development of product specifications and data delivery models); and
- develop and test approaches for gaining an improved understanding of Australian government policy needs related to biodiversity profiling (framed by the feasibility of their implementation).

2.3 Assumptions and caveats

In conducting this study, it was assumed that the study:

- would focus on prototypes and not aim to develop a functioning and operational biodiversity profiling capability for Australia; and
- would not undertake significant new research but rather focus on achieving innovation through the coupling of ecological data, satellite-derived metrics of landscape change (such as vegetation condition) and novel modelling techniques.
- The study also assumes that there is a requirement for spatially explicit measures of change in the state of biodiversity at continental scales.

2.4 Organising framework and report outline

Table 1 shows the organising framework that was developed at the start of the study to scope key components of developing an operational capability. It is not necessarily comprehensive but provides a high-level overview of the key components that received treatment in this study. The organising framework was traversed by a series of work packages that translate into report sections.

- **Section 3: Enabling accessible data** describes documentation of Information Communications Technology (ICT), licensing and governance enablers for federating natural history collection data, and documentation of knowledge built in the development of the ALA. Although findings and recommendations are based on ALA experiences, many of the issues are generic and are likely to be useful for other activities that take a federated approach to information system development.
- **Section 4: Atlas of Living Australia (ALA) Data and Tools** covers exploratory data analysis documenting taxonomic, spatial and temporal extents, currency and data types. It provides a summary of the ALA data and tools (reporting) that have emerged from the implementation of the processes documented in Section 3 above.
- **Section 5: Business requirements elicitation** describes the development and testing of an approach for assessing Australian Government policy needs framed by feasibility as informed by ALA and CSIRO experts and capability. Lessons from this are partially used to inform the type of analysis conducted in the later components of the study, and in particular, the modelling component.
- **Section 6: Case study – biodiversity profiling** provides an example that integrates selected ALA data and ancillary spatial information to make inferences of biodiversity change at continental scales. The section also briefly examines the availability of other supporting national spatial environmental data (such as land use and land cover) that could be coupled with ALA observations to establish an operational monitoring capability.
- **Section 7: Conclusion** provides a summary of key lessons and closing recommendations that emerged from the study.

Table 1. Study organising framework showing key components.

Components					
	Enablers		Data	Reporting and data delivery	Analysis and modelling
	Soft	Hard			
Elements	Licensing and governance levels, (technical and policy steering committees), resourcing, negotiating and resourcing third-party data access, maintenance agreements, IP arrangements, legal liability and risk management, citizen science models.	Information systems, web data services, architectures, interoperability, semantic web, standards, field-based computing, data transfer standards.	Ecological, contextual (abiotic), time series satellite imagery.	Data services, dashboards, visualisation technologies, reporting tools.	Algorithms, models, research to develop new approaches for integrating data, model data fusion.
Examples	Creative Commons, AusGoal, ALA models for licensing, Birds Australia Bird Atlas, citizen science and community engagement programs.	Open Geospatial Consortium, ISO19115 Metadata, Darwin Core, TDWG.	ALA, SEWPaC, TERN, Birds Australia, Covariates (GeoFabric, climate, land use and land cover), NCAS vegetation mapping, satellite imagery, NVIS mapping.	OGC Web Feature Service, OPeNDAP, NetCDF, mapping applications, dashboards, report cards.	Generalised Dissimilarity Modelling, Generalised Linear Modelling, image processing, data assimilation, prediction and inference.
					Stakeholder workshops; meta analysis and synthesis; multi-criteria evaluation.
					Methods for user needs assessment, identifying business drivers, developing functional specifications, product specifications, models for engagement.
					Historical trend detection, State of the Environment reporting, climate change impacts (national reserve system planning), environmental accounting, program performance monitoring (e.g. Caring for Our Country, Carbon Farming Initiative, Biodiversity Fund).

3. Enabling accessible data

This section explores the technical and communications aspects involved in facilitating access to the widely distributed data available on Australia's ecology. Many of the concepts are applicable to other domains. The discussion and recommendations offered are based on the experiences of the Atlas of Living Australia in aggregating national biodiversity data.

3.1 Background: the Atlas of Living Australia

The mission of the Atlas of Living Australia (ALA) is 'to develop an authoritative, freely accessible, distributed and federated biodiversity data management system for the Australian region'. The Atlas was initiated by a group of 14 (now 17) organisations. The intent was to create a national database of all of Australia's flora and fauna that could be accessed through a single, easy-to-use website. Information on the website would be used to:

- improve our understanding of Australia's biodiversity;
- assist scientists to build a more detailed picture of Australia's biodiversity; and
- assist environmental managers and policymakers develop more effective means of managing and sustaining Australia's biodiversity.

The ALA combines biological and environmental data with a suite of tools (Appendix 1) and a web interface to demonstrate the utility of the project. The features of the ALA can be explored through its web portal (<http://www.ala.org.au>, last accessed 1 May 2013), and most of its functionality can be obtained through web services. This strategy has enabled rapid development of new portals for the Australian Virtual Herbarium and the Online Zoological Collections of Australian Museums using ALA web services. Such services readily enable any existing website to tap most of the power of the Atlas.

The current partners of the ALA include:

- CSIRO—as the administrative lead agency
- Australian Museum (AM)
- Museum and Art Gallery of Northern Territory (MAGNT)
- Museum Victoria (MV)
- Queensland Museum (QM)
- South Australian Museum (SAMA)
- Tasmanian Museum and Art Gallery (TMAG)
- Western Australian Museum (WAM)
- Council of Australasian Museum Directors (CAMD)
- Council of Heads of Australasian Herbaria (CHAH)
- Council of Heads of Australian Collections of Microorganisms (CHACM)
- Council of Heads of Australian Entomological Collections (CHAECE)
- Council of Heads of Australian Faunal Collections (CHAFC)
- Southern Cross University
- The University of Adelaide
- The Australian Government Department of Agriculture, Fisheries and Forestry (DAFF)
- The Australian Government Department of Sustainability, Environment, Water, Population and Communities (DSEWPaC).

In developing the Atlas, strong relationships were created for sharing biodiversity data, resources and experiences with related international projects such as:

- The Encyclopaedia of Life (EOL)
- Biodiversity Heritage Library (BHL)
- European Union's Distributed Dynamic Diversity Databases for Life (4D4Life)
- Data Observation Network for Earth (DataONE)
- Barcode of Life Database (BOLD)
- Global Biodiversity Information Facility (GBIF)—the Atlas is the Australian node of GBIF

- Map of Life (MoL)
- The EU's CReATIVE-B project.

The Atlas has also worked closely with other NCRIS funded projects to develop common approaches and standards and to share application development where possible:

- Terrestrial Ecosystem Research Network (TERN)
- Integrated Marine Observing System (IMOS)
- Australian Biosecurity Intelligence Network (ABIN)
- Australian Phenomics Network (APN)
- Australian Plant Phenomics Facility (APPF)
- Australian National Data Service (ANDS).

The Atlas of Living Australia has delivered a national biodiversity data warehouse management system which links Australia's biological knowledge with its scientific and agricultural reference collections and with other custodians of biological information. A key outcome of the ALA has been to successfully integrate information on Australian species, including data on specimens held by Australia's natural history collections and data from field observations of living organisms. As a measure of this success, the ALA has over 35 million species occurrence records which can be explored and analysed in the context of over 380 environmental layers (see <http://dashboard.ala.org.au>).

As an infrastructure project, the ALA does not generate primary data itself, and so to fulfill its mission and meet the project aims, the ALA identifies and aggregates biodiversity data generated and managed by diverse sources distributed across Australia and the world.

The Atlas of Living Australia's successes in aggregating biodiversity data have been partly due to two key principles—supporting the sources and custodians of data and making the aggregated data as highly usable as possible through the addition of data discovery, visualisation and analysis tools. Biodiversity data aggregated by the ALA show a

consistently increasing volume of records being collected and a significant jump in the amount of data collected in the last decade. The number of records doubled in 2000–2009 compared to 1990–1999.

The ALA has so far only aggregated the obvious and relatively easily accessible sources, many of them aggregators themselves. There is significantly more data available, and this will be in many smaller and highly distributed datasets. Much valuable biodiversity data has however already been lost. It is therefore a priority for aggregators such as the Atlas of Living Australia to identify and mobilise as many of the remaining datasets as possible and as quickly as possible.

3.2 Data aggregation

The aim of aggregating data is to provide a single, standardised mechanism of access to data sourced from many, often distributed and heterogeneous, data stores. Ecological data are generated from a multitude of activities carried out by professionals and amateurs as individuals or a part of an organisation that includes government departments, community interest groups, research institutions and commercial companies. The number of distinct datasets generated by these activities and the volume of records needed for analysis renders a distributed access mechanism as impractical. An aggregated view of the data via a central system is the only current feasible solution. A centralised data store assembled this way is known as a data warehouse.

3.2.1 A warehouse of ecological data

The process of warehousing ecological data focuses on the aggregation of records already captured in existing databases, spreadsheets and documents (some not digitised). Much of the existing data was not collected with standards, metadata and sharing (or possibly even digitisation for some of the older

sources), aggregation or large-scale analysis in mind. These data have been collected, integrated and stored for ease of collection management, local conservation and research purposes. The data are in varied structures across sources, different information is held or different terms are used for the same information. Where information has been aggregated it may be incomplete within a single source, for example, a database compiled from different datasets or entered by different people each with a different use of 'name' field. The data are on a variety of systems and platforms and often lack technical information (such as database and transformation descriptions) and context metadata (when, where, how and why the data were collected).

Ideally, these data should be managed and reconciled at the source, standardised and have metadata generated. The data can then be more easily mobilised and aggregated to form the foundation store of biodiversity information. Management and reconciliation of data is an intensive task requiring domain specific knowledge. The primary benefit (aside from the information being digitised) of the effort involved in completing, documenting, mobilising and aggregating these data is that it removes the need for multiple organisations to perform this task over and over. In practice, there are many blockers to this model in that the original data providers may have moved on and/or resources (both human and financial) are simply not available to enable the at-source processing of these datasets. The aggregator can perform an effective role in educating data providers about standards and processes.

3.2.2 Facilitating 'shareable' new data

Although warehousing efforts concentrate on existing (sometimes referred to as 'legacy') data, it is also necessary to be aware of the volume of data that are being collected now and will be collected in the future. Encouraging communities to capture data

electronically, the use of standards and effective data management processes will make the aggregation of this information much more efficient in the future.

Through continued natural resource management, research and citizen science collection efforts, there are infinite biodiversity-related data to be collected in the future. There is an opportunity to influence the standardisation of data capture and storage before an increasing variety of methods, standards and technologies proliferate. The adoption of best-practice data models, collection practices, stable identifiers, embedded metadata, quality and sharing capabilities into existing and proposed collection systems are required for efficient investment in the domain. This work is facilitated by providing free, easy to use systems that are standards compliant, mobilisation ready, and that promote quality and maintain lineage.

There are now strong legislative and licensing frameworks to ensure the sharing of publicly-funded research and other data (Declaration of Open Government, AusGOAL and Creative Commons). The next step is to provide a legislative framework and associated policies that mandate the data to be captured during relevant activities and, if available, to ensure that these data are provided to the relevant national facilities.

3.3 Key principles

The approach of the ALA in aggregating data has been guided by two key principles: support the interests of primary data sources and custodians and provide additional capabilities on the aggregated data.

3.3.1 Support the interests of primary data sources and custodians

It is vital not to diminish the resources (i.e. funding) available to collect primary sources of data. A pure aggregator does not generate data and is totally

dependent on primary data sources. The more those sources are able to engage in their core activities and collect data, the more that these data can be made available. Aggregators should at least assist custodians by providing useful data-sharing infrastructure, so that they can direct their resources to focus on their core activities rather than resolving IT problems. The ALA supports data sources and custodians using the following actions.

- Ensuring proper recognition of the source of the data and prominently displaying licensing and attribution for the shared datasets.
- Where data are maintained at the source facilitating the feedback of annotations and additions to the primary source. New and edited records should come via the source rather than records being edited in the aggregated environment.
- Minimising the technical and administrative burden on sources and custodians.
- Using patience and understanding issues in the primary source's domain in the absence of policy or legislative powers to obtain data.
- Providing clear and understandable advice in relation to data sharing issues (particularly licensing).

3.3.2 Provide additional capabilities on the aggregated data

Although there is value in aggregated data simply in saving users time in identifying, contacting and arranging access through multiple data providers, if it is to be used in analysis and decision support, there need to be additional services provided. The following should be included.

- Building trust and understanding in the data itself via the central repository. In this respect, it is important to note that the ALA itself does not actively 'vet' or filter data. The ALA's approach is to make available as much metadata as possible, including the results of 'data quality' checks. This

approach puts tools in the hands of users so they can effectively determine which data are fit for their particular needs and filter it to reach that subset.

- Sharing usable information between aggregators through transparent, traceable and complete data.
 - Where derived or standardised values are added to source data, keep the original values and log the changes.
 - Maintain keys to the source system so the original records can be identified.
 - Encourage and facilitate the sharing of complete records wherever possible, including verbatim values, precision and uncertainty and other supporting fields (information on the who, what and when of the records).
 - Use standards or defined fields from standards wherever possible so definitions are clear.
 - For flexibility, standardise the central repository across a limited set of core elements but allow the inclusion of almost any additional information as extensions.
- Providing interfaces tailored to different communities of use. For example, the ALA provides both a simple interface that can be used for rapid assessment of species or areas and an advanced interface for extensive and open-ended analyses.
- Allowing for simple upload of *ad hoc* data to be combined with aggregated data, and for downloading data in most usable formats.
- Providing tools that demonstrate the significant value added by integrated biological and environmental data.

3.4 Including a resource in a data warehouse

There are three high-level steps to including resources in a data warehouse:

- identifying potential sources of information to share and contacting them to discuss sharing their data;
- moving data to accessible infrastructure; and
- aggregating the data into the central repository.

3.4.1 Identification and communication

There are many types of entities operating in the generation, management and use of biodiversity information. An aggregator interacts with all of them. Biodiversity data have been generated by thousands of activities every year by researchers, natural resource management agencies, community and conservation groups and interested individuals. Stakeholders can be loosely categorised as primary sources and aggregators although some organisations may generate their own data and aggregate others' data (often the case with government agencies).

Other organisations, infrastructure projects and research projects also aggregate data—at varying scales (e.g. global, national, State and local) and scopes (e.g. birds, all biodiversity, vegetation or marine). State and federal agencies aggregate data within their jurisdictions to support monitoring of, reporting on and management of their biodiversity, environment and natural resources. Government agencies often generate their own primary data as well through surveys and monitoring activities.

The following should be considered when engaging with aggregators.

- The information management area that aggregates the information within the organisation is not necessarily the 'owner' (that can give permission to share the data) and this may be a barrier to sharing the data.

- Aggregated data have come from many sources so awareness of the sources that have contributed to the dataset will help to reduce duplication. For example, State conservation agencies aggregate data from other entities including their own surveys, private organisations, museums and herbaria.
- Aggregation initiatives may use different licensing models and incentives which may set particular expectations or provide information on data sharing principles and practices that can support or run counter to those the project supports.
- Collaboration with other aggregation initiatives can reduce over-communication with primary sources.
- Each project will have its own priorities and management may be problematic.
- If it is possible to establish common licensing frameworks and principles across aggregation initiatives this will create an expectation in the community as to the 'standard operating procedure' for sharing biodiversity data. This will greatly simplify access and licensing negotiation for all aggregators. For example, the ALA has generally adopted the Creative Commons suite of licenses and recommends the use of CC-BY wherever possible.
- It may be more efficient to access data via aggregators but there are often additional levels of management and administrative processes to be negotiated and the data will have been processed to meet their needs rather than more generic requirements.

A concern raised by data custodians (particularly aggregators but also individual providers) relates to not having the permission of their contributors to share the data. To get permission from all the contributors to a dataset may be difficult. It is important to recognise the rights of contributors, and aside from the legal implications, releasing their work without permission can easily have a negative effect on their willingness to contribute in the future. If it is not possible to contact the owners, then the

information may be able to be released with a ‘take down’ notice indicating that if this information has been inappropriately released, it will be taken down immediately if contacted. For more information on licensing see *3.4.1.6 Legal matters – Rights, Licensing and Terms of Use*. This issue highlights the need for sharing data to be a consideration at the point where it is collected and managed.

Primary biodiversity occurrence data are generated through surveys, ad hoc or incidental sightings and specimen collection activities carried out by scientists, conservation groups and government and non-government agencies and individuals.

The following should be considered when engaging with primary sources.

- Although accessing data directly from the source provides the original records, these data are in a multitude of smaller datasets each with an analysis and transformation overhead to prepare it for sharing. It may be the same amount of effort to prepare a dataset of 100 records as one of 1,000,000 records.
- Individuals and local conservation groups are, in general, very open to sharing their data as their goals are to increase general awareness of the biodiversity of their region and their own profile.
- Scientists will usually need to keep their data private until they publish the results of their research. Given that research data are usually aggregated by a scientist over many years, there is often a reluctance to release the data because it is incomplete and/or there is an investment against future publications. Note however, that metadata can be made public at the outset of a project.
- Primary sources may donate their data to one or more aggregators, for example, bird records are often in several datasets including an original, a local birding group, Birdlife Australia and a State agency. De-duplication by the aggregator can largely address this issue.

- Owners of data have varying amounts of time they can devote to sharing their data. Some are happy to devote significant time to collecting together, digitising where needed and reviewing and cleaning up their records in preparation for sharing. Others are happy to share but cannot put in any time to standardise their data or install mobilisation systems.

As the creator of the records, primary sources usually (but not always) are able to grant permission to share their data. Situations where they are not able to grant permission are where they do not own the copyright—usually this is the case for data generated while employed. This can cause some conflict when an individual wishes to share their work and the organisation they work(ed) for is reluctant.

3.4.1.1 Identifying resources

The large aggregators of information are usually well-known in the field, and in biodiversity they are the State conservation agencies, Birds Australia and the State and national natural history collections. Other sources can be identified by reviewing the sources of the data in the aggregates, contacting research and education institutions. Even broad internet searches are useful to find local conservation groups. Examples include:

- central State and Federal government agencies;
- scientific organisations;
- natural history collections;
- regional management, councils, catchment management authorities;
- natural resource management groups;
- volunteer and other conservation agencies;
- societies, and interest groups (fungi, birds, herpetological, plants, parks etc.); and
- domain specialists/individuals.

Once an aggregation initiative such as the Atlas of Living Australia is established and providing value, the community may make contact with offers of data to be shared. This is a positive indicator of the state of engagement of the project with the community of data owners and custodians. At this stage the aggregator can begin to move some of the burden of mobilisation to the owners of the data through requesting the data be set-up or supplied in the standard format or run through a pre-qualification process before submission. Pre-processed or verified data greatly speed up the integration of datasets. This work will be necessary if the smaller datasets are to be loaded as a large number of smaller datasets takes much longer to review, map and transform than a single large dataset.

3.4.1.2 General advice

The ALA has managed communications with potential contributors by maintaining a light touch, noting that sending out reminders and follow-ups without being perceived as ‘pushy’ and putting contributors offside is a challenge. This approach has been generally successful and the volume of data shared via the ALA has therefore continued to expand. Due to the nature of the project to develop infrastructure, the ALA was not able to provide direct financial incentives for sharing data. Such an arrangement would be unsupportable over time for an aggregation initiative unless they, in turn, charge for access. With this in mind, sharing data can be looked at as similar to volunteering or donating to a charity or community event.

Potential contributors that do not engage following the initial ‘soft’ approach are a challenge to work with to mobilise their data. It is not practical to enforce data sharing even where an agreement or other instrument exists. In these cases the data may be old or of poor quality. It may be necessary to withdraw from an engagement to focus on others more willing or able to engage. After a suitable interval, the provider may be re-contacted

to reassess their situation. If this continues to be unsuccessful or the resource is of particular importance other avenues to encourage sharing the data include:

- seeking requests from users or contributors to their data to request to share it;
- seeking executive support to share data from within their organisation;
- referring to government guidelines and policy (e.g. open government);
- seeking direction from funding bodies (e.g. government); and
- seeking support from other initiatives (Creative Commons Australia, AusGOAL).

Contact with current and potential information suppliers needs to be co-ordinated and managed to ensure that a correct and consistent message is presented. Co-ordinated communication at executive management level and operational levels promotes the advantages of sharing information and encourages effective co-operation between data owners and the aggregator. In practice, the ALA has found that building strong relationships with the individuals at the operational level is vital.

It is important to recognise that those involved in data sharing will also have other responsibilities that may be a higher priority. It often takes several months for a data custodian to become comfortable with the concept of sharing their data, to consider and select terms of use and prepare and transfer their data. It is usual for administrative issues to take considerably longer than dealing with the technical issues.

There may be other management or oversight structures that are stakeholders in the data managed by custodians. Examples of such groups include the Council of Heads of Australasian Herbaria (CHAH) and the Council of Heads of Australian Faunal Collections (CHAFC). There are similar groups for entomology collections, botanical gardens and

microorganisms. These groups may have legal rights over data or have a role as a reference group. It is therefore important to include them in communication with their members.

Coordination with related projects and institutions to form a consortium for information sharing may give the information-sharing community a consolidated approach to larger institutions. This will also reduce the likelihood of over-communication (that is, multiple requests to share data). Realistically, unless the projects are collaborating towards the same deliverables, not the same type of deliverable, but a single deliverable jointly delivered by both projects, each project will be working with their own timelines and resources to meet their own goals. Without a common goal, the divergence tends to make this level of coordination impractical, beyond raising awareness of the other initiatives.

3.4.1.3 Initial contact

One key aim of initial contact is to build a relationship with an individual that can be a 'champion' within the organisation, strong personal relationships will ultimately drive the success of collaborations in sharing data. The champion can either provide permission and access the data themselves or manage the process within their organisation to gain permission and access. Managing multiple relationships within the organisation is much more difficult and time consuming for an external party. The ALA experience highlights the need to approach potential contributors via an introduction whenever possible. An introduction from a trusted or recognised colleague or individual within the same field immediately gives the contact and project greater credibility. Initial contact at the operational level (those that have day-to-day responsibility for managing and providing access to data) has worked well for the ALA. If executive permission is needed then the person in the organisation has either managed that themselves or provided the relevant contact.

Advice for developing an initial contact message includes:

- Keep the message brief and to the point.
- Introduce the project, include examples of other similar and well-known groups or initiatives that are sharing their data. Knowing that the State conservation agency and other groups with the same focus are involved gives immediate credibility.
- Indicate a few simple and obvious benefits (for example, answer the initial question 'what's in it for me?').
- Mention that it does not cost anything (that is, answer the initial question 'what will this cost?') but that some of their time will be required.
- End with thank you and a question. Information on your project is important but the potential contributors need to think about what, if and how they may be able to contribute. A specific question also helps to encourage a response. For example, 'Is sharing your data something you would be comfortable with?' is more likely to elicit a response than 'Please let us know if you would be interested in contributing'.

Following an initial contact, very few will respond quickly (either positively or negatively) and most will not respond at all. It is recommended that potential contributors are given at least a few days before following up.

3.4.1.4 Motivation and barriers to sharing data

Once a data custodian has been introduced to the project and potential for sharing their data, a series of contacts will usually follow. These may be face-to-face (individual meetings, presentations or workshops) or by phone or email and can discuss the motivations to share data, any perceived and actual barriers to be overcome and the process of mobilising data. Face-to-face contacts are best, but Skype or equivalent may be a useful option if that is not possible. Basic questions such as 'why should

I share my data?’ or more accurately ‘what’s in it for me?’ need to be addressed to justify the work necessary in sharing data.

Some possible motivators for sharing data may include:

- the more information that is available, the better informed decisions will be;
- assisting in conservation of a species or area by raising awareness;
- social aspects (some people simply like to share their photos, knowledge and passion);
- raising awareness of their personal or organisations activities (for recognition or opportunities for collaboration or funded projects);
- a prerequisite of funding, some funding sources insist that the data are published or deposited in a public data store;
- open government principles; and
- aligning with their organisation’s mission (for example, ‘The Museum’s mission is to increase understanding of, and influence public debate on, the natural environment, human societies and human interaction with the environment’).

3.4.1.5 Contributor prioritisation

Identifying potential sources of data will depend on the objectives and mission of the data provider and a choice of appropriateness for aggregation will be influenced by the following:

- existing partnership agreements;
- a geographic, temporal or taxonomic focus;
- number of records;
- data quality (defined in the context of its collection);
- ease of mobilisation (for those that have facilities already in place); and
- ease of licensing (for those that already have open licensing frameworks in place).

3.4.1.6 Legal matters—rights, licensing and terms of use

Facilitating general access to an aggregated dataset requires the permission of the creator, custodian or owner of the source datasets (the ‘rights holder’). The permissions required relate to both the legal rights that owners have over their work and cooperation needed from custodians to gain access to the data. Essentially there are intellectual property law and/or contract law considerations to be resolved to share material created and owned by others. When aggregating data from a number of sources it is important to acknowledge their contributions and ensure that the conditions under which the information is made available are adhered to. The form of the acknowledgment and the terms of use are managed in a license.

Any discussion of legal matters is intended to highlight and discuss a common issue and is not a comprehensive exploration of all the potential legal issues. Specific legal advice needs to be obtained for particular circumstances.

3.4.1.6.1 Intellectual Property and copyright

(From: <http://www.ag.gov.au/Copyright/Pages/Whatisintellectualproperty.aspx>, © Commonwealth of Australia 2011. Used under a [Creative Commons Attribution 3.0 Australia](#) license).

‘Intellectual property law protects the property rights in creative and inventive endeavours and gives creators and inventors certain exclusive economic rights, generally for a limited time, to deal with their creative works or inventions. This legal protection is designed as a reward to creators to encourage further intellectual creativity and innovation, as well as enabling access by the community to the products of intellectual property. Because intellectual property protects rights, rather than physical property, intellectual property is an intangible form of property. It is property which cannot be seen or touched.

Intellectual property is the general name given to the laws covering patents, trade marks, designs, circuit layouts, plant breeder's rights and copyright. Each of these forms of intellectual property is protected by a specific Act of the Commonwealth Parliament. The framework for these Acts is largely based on Australia's obligations under international treaties.'

(From: <http://www.ag.gov.au/Copyright/Pages/Whatiscopyright.aspx>, © Commonwealth of Australia 2011. Used under a [Creative Commons Attribution 3.0 Australia](#) license).

'Copyright is a type of property that is founded on a person's creative skill and labour. It is designed to prevent the unauthorised use by others of a work, that is, the original form in which an idea or information has been expressed by the creator.'

Copyright is not a tangible thing. It is made up of a bundle of exclusive economic rights to do certain acts with an original work or other copyright subject-matter. These rights include the right to copy, publish, communicate (eg broadcast, make available online) and publicly perform the copyright material.

*Copyright creators also have a number of non-economic rights. These are known as moral rights. This term derives from the French *droit moral*. Moral rights recognised in Australia are the right of integrity of authorship, the right of attribution of authorship and the right against false attribution of authorship.'*

Copyright comes into existence automatically at the time a work is created and does not require registration (in Australia). Of particular relevance to data aggregators is the question of whether copyright exists in data. To be able use a license that relies on copyright (this includes the Creative Commons Licenses discussed below) copyright must exist in the 'work'. Several court cases both in

Australia and overseas have both upheld and denied the existence of copyright in specific databases and there is some excellent information discussing this issue on the Australian National Data Service website (<http://www.ands.org.au/guides/copyright-and-data-awareness.html>) including:

Copyright applies to text, images, recordings, videos etc in a digital form, in the same way that it would to analogue versions of those works. The code of computer programs (both the human readable source code and the machine readable object code) is protected by copyright as a literary work. Data compilations such as datasets and databases can be protected by copyright in the literary works category, which includes 'tables' or 'compilations'. A table or compilation, consisting of words, figures or symbols (or a combination of these) is protected if it is a literary work and has the required degree of originality.

Mere information or a random collection and listing of unrelated facts or data will not be considered to be a compilation for copyright purposes. However, a factual compilation will be a literary work if it supplies 'intelligible information'. It will be protected by copyright as an original literary work if it has been produced by the application of independent intellectual effort by the author/s, which may involve the exercise of skill, judgment, knowledge, creativity or labour in selecting, presenting or arranging the information. Copyright applies not to the facts/information itself, but to the particular way the facts/information are presented in the dataset or database.

(From: <http://www.ag.gov.au/Copyright/Pages/CancopyrightbeboughtandsoldinAustralia.aspx>, © Commonwealth of Australia 2011. Used under a [Creative Commons Attribution 3.0 Australia](#) license)

'Copyright can be dealt with in the same way as other forms of personal property. It can be assigned, licensed, given away, sold, left by will, or passed on according to the laws relating to intestacy or bankruptcy. This does not apply

to moral rights which are personal and which creators cannot transfer or assign. It is always best to obtain written evidence of permission to use copyright, rather than rely on oral statements.

Assignment

An assignment of copyright must be in writing and signed by or on behalf of the assignor (ie the copyright owner) to be legally effective. The assignment may be in whole or part and may be limited to one or more of the exclusive rights or aspects of them and may also be limited as to time or geographical area.

Exclusive licence

An exclusive licence grants specified rights to the licensee with a guarantee that those rights will be granted to no other person. An exclusive licensee can sue and take certain other actions as though he or she were the copyright owner. Exclusive licences, like assignments, must be in writing and signed.

Non-exclusive licence

A non-exclusive licence is a permission to exercise one or more of the copyright owner's rights in a work. It does not result in the copyright owner parting with his or her rights in the work. A copyright owner may grant numerous non-exclusive licences, but can assign any or all of the exclusive rights that comprise his or her copyright only once for the period of that assignment.'

Note that a license grants others permission to exercise specific rights but does not result in the copyright owner 'giving away' those rights permanently. This is an important distinction to make for people considering sharing their data that are concerned about a loss of ownership.

3.4.1.6.2 Licensing

Information that does not fall under copyright is not necessarily freely available and an aggregator may need a contractual arrangement to gain access.

Different license types can apply to different types of information; some materials (e.g. raw facts) are not covered by copyright. A contractual license is essentially a contract between legal entities granting specific rights while a copyright based license relies on the automatic existence of copyright in original works. A license can only be granted by a person that owns the rights to the work. Who owns the rights in a resource (and so can pass those rights on) is not necessarily the same as the person that currently has custody of the resource or even the person that authored the resource. When obtaining permission to share information, it is important to ensure that the person granting permission actually has the rights to do so.

An aggregator needs to have permission to share the resources that they collect with others. This permission will either involve sub-licensing or a direct license and may be contractual or copyright-based. A license only granting rights for the aggregator to use the work, but not distribute it or derivatives of it, will effectively cancel out the benefits of having it in the aggregated dataset.

When aggregating data to be shared with others, open access licensing allows owners of resources to share information broadly, not just with individuals one at a time via a contractual type data of use agreement. Due to the different types of information the ALA aggregates (including data, images, movies and text) Creative Commons was selected as the preferred licensing method.

Creative Commons provides simplicity in implementation—all that is required to release a work under a Creative Commons license is to label the work appropriately (there are no signatures required). Creative Commons also provides a common grant of rights applicable across different types of works (data, images, film, text, etc.) which is highly useful for aggregators of a variety of information types.

Efficient sharing of data is very difficult unless a common licensing framework is adopted for access and re-use of data. It is the ability to combine data from different datasets and contributors that really provides value to the users of the aggregates. Creative Commons licences allow the ALA to require users to attribute the source of data and to restrict use to non-commercial applications. There are licenses in development such as Open Data Commons that may prove to be more suitable for the ALA but until a more appropriate license becomes available the ALA will continue to encourage the use of Creative Commons Australia licensing. Creative Commons Attribution 3.0 Australia licence allows users to copy, transmit and reuse the information, and to remix or adapt the information as long as attribution regarding the source of the information is maintained.

Additional conditions may optionally be added to the basic Creative Commons Attribution licence from those listed below.

- **Non-commercial:** Users may not use data from the supplier for commercial purposes.
- **Share alike:** If users alter, transform, or build upon the data from the supplier, they may distribute the resulting work only under the same or similar licence to the licence selected for the data resource.
- **No derivatives:** The user may redistribute the work in unaltered form only. The ALA does not aggregate data (although it may be appropriate for a document) under this license as it does not allow the level of reuse that the project aims to provide.

3.4.1.6.3 ALA data provider agreement

The release of information under a Creative Commons license is sufficient for the ALA to be able to aggregate and share. However, for situations where the ALA established a longer-term relationship with a data provider and/or the data provider wished a more formalised relationship, the ALA developed a Data Provider Agreement.

The Data Provider Agreement is a statement from a contributor of the terms under which they wish to share their information via the ALA. The terms of use includes intellectual property rights and assurances from the contributor that they will maintain the information and have the rights to share it. A core component is the conditions under which others may use the information and ownership of any derivatives of the information. The agreement document allows for the selection of Creative Commons licensing for this purpose. It is also possible to create a custom license although this not preferred. The key terms of the data provider agreement are listed below.

Contributors rights and responsibilities

- Contributors retain IP rights and ownership of their data.
- Contributors are responsible for either managing any data restrictions (providing the data in a de-sensitised form) or informing the ALA of any data restrictions they require the ALA to implement, for example, locations of threatened species (the ALA may implement management if they aware of the rules and actions).
- Contributors endeavour to ensure the data are accurate.
- Contributors assert that they have the rights to provide the data (they own it).
- Contributors are responsible for updating and maintaining their data.
- Data will be made available to the ALA free of charge.
- Data will be made available under a Creative Commons license selected by the owner.

ALA use of the data

- Data will be attributed to its contributor.
- Information may be made available via the internet in both raw (individual records) and aggregated or derived form (e.g. a density map or count of species).

- The ALA may cache (keep a copy of) the data.
- The ALA may pass data on to international partners (e.g. GBIF, EOL).
- Contributors may be given a rating of the quality of their data.
- The ALA will remove the data from publication if requested by the owner.

Information use agreement

The terms of use for the ALA systems includes an information use agreement. The terms include general provisions and that there are specific terms of use according to the license selected by the data provider. By using the ALA, a user agrees to the terms of use (<http://www.ala.org.au/about-the-atlas/terms-of-use/>).

3.4.2 Mobilisation

Mobilising data involves retrieving the records from the source system, converting them to a standard format and moving them to the central repository. The mobilisation steps and licensing discussion are not linearly dependent on each other but all of them need to be completed before the data can be loaded, processed and made available. Some common scenarios include:

- A sample of data are provided so the owner can see how it will be made available before they commit to sharing the entire dataset.
- The entire dataset is provided for analysis and mapping while the owner considers the terms of use they would like to apply.
- The owner of the data exports, maps and transforms the data for sharing and once they have submitted the data become aware they need to set terms of use and attribution.

Although the aim of mobilisation is to have individual records available to be accessed as part of the aggregated dataset, the following other levels of mobilisation still provide value.

- **Discovery:** the resource is able to be found, this requires metadata or information about the resource and what it contains. An example of this is a metadata entry describing a database of bird observations where the database is either not online or not publicly available. A search on the relevant species will identify the database as a potential source of information.
- **Resource level access:** the resource can be accessed but the individual records or units of information within it are not available as part of the aggregated dataset. An example of this is a scanned monograph containing information on the distribution of species that is available on line. The monograph can be found as relating to the species and read but the sightings recorded in the document are not available as records in the dataset
- **Record level access:** the individual records or units of information within the resource can be individually accessed as a seamless part of the aggregation.

3.4.2.1 Mobilisation architectures

There are two basic architectural models that can be followed to create an aggregated dataset including (a) distributed architectures that create a virtual interface using remote access, and (b) centralising the data (warehousing).

3.4.2.1.1 Distributed architectures

Distributed architectures function by accepting a query at a central location in a common syntax and then sending that query on to the interface built for each of the separate source databases, the individual data sources run the query and return any results in a standard format. Standard interface to all the source databases to accept a remote query and return standardised data. These systems are moderately complex to install, configure and maintain (they require solid IT knowledge and experience). In theory the data is always current as it is being retrieved in real time from the source

database; in practice remote access is usually provided to a copy of the live data to separate the processing load from the operational system and address security concerns. This cancels out the main advantage of this architecture model. Infrastructure needed by the aggregator doesn't have to maintain as much permanent storage as that needed for a warehouse.

The bandwidth needed and response times are impractical as the size of the system and results sets scale up. For example, it can take several days to return a large (million record) result but result sets of this size and larger are common for large-scale analysis. So although systems designed for distributed architectures can be used to centralise data they are inefficient for large volumes.

A distributed architecture does not operate well for many smaller static datasets like spreadsheets. The overhead and infrastructure is the same for a few hundred records in a spreadsheet as for hundreds of thousands of records in a managed database. Distributed architectures are a more effective solution for aggregating a few smaller, highly active source databases (within a single infrastructure if possible to avoid security problems) where data currency is a high priority.

3.4.2.1.2 Centralised architectures

A centralised architecture maintains a single data store (data warehouse). Data are exported from each of the distributed source databases, transferred to the site of the warehouse and loaded into the single standardised database. Queries are run against the compiled dataset. The following are key components and characteristics.

- Mapping and transformation to the standard format can be performed at the source or the aggregate depending on where the knowledge and resources are available.
- Source systems do not need to all have the same infrastructure installed to share data. Export and

transfer systems can be very simple using the existing data management technology which data owners are comfortable and knowledgeable in. This leaves data owners able to focus on their work rather than IT installation and maintenance. This also makes centralised architectures relatively efficient for smaller static datasets like spreadsheets.

- As data are exported, transferred and loaded, there is a lag before data become available in the aggregate. The time to load a dataset depends on whether it is in a standard format when it is received and if not, how much mapping and transformation is needed to get the data into a standard format, including how engaged the data owner is in helping with the mapping. For repeated loads, as long as the format is stable then transformation processing can be re-used.
- The aggregator needs to maintain significant infrastructure to store and manage the data in the data warehouse.
- Response times are much faster particularly for large queries making this a much more effective architecture for interactive investigation of the data.
- It is much more efficient to transfer large volumes of data in simple text files than as streams of marked up text. This makes initial loads and updates of larger data bases much easier and significantly more stable. Exporting and sending files is also not a significant security concern.
- Centralised architectures are suited to aggregating large numbers of distributed data sources of varying sizes and structures.

3.4.2.2 Mobilising data

The mobilisation strategy follows a standard pattern but adapts to individual data source environments. This results in common approaches but with the detailed solution being tailored on an institution basis, that is, based on a 'best fit' solution for the architecture. This takes into account individual data

management practices, systems, workflows and the IT context including overall technical maturity and network (including firewall issues). Mobilisation aims to adhere to the following principles:

- If there is an effective, efficient mobilisation system in place then improve the completeness of the data feed if needed but do not replace the system without good reason.
- The source system owners must be able to maintain the mobilisation solution (also following the need to keep it simple).
- Darwin Core (see <http://rs.tdwg.org/dwc/>) as the target exchange schema is the point of commonality.
- Move files around rather than allow remote querying to ensure greater stability of the solution and tighter integration.
- Delimited text files are the preferred format (Darwin Core Archive).
- Re-use components where the infrastructure is common (e.g. KEmu systems).

Where no other solution exists, the simplest approach typically involves an export from the source system to a flat Darwin core csv file. This file is then sent via FTP to the ALA server where it is loaded into the ALA database with the help of a number of scripts. Once the file structure is stable, the receipt and load can be automatically triggered.

Individual data stores differ mainly in the development and implementation of the export query, the remainder of the process is standard (for the platform). This is a 'push' process rather than 'pull'. This type of transfer keeps the process simple and requires far less security modifications (if any) to the source systems. The process has a number of steps that are highly suited to automation:

- re-usable export query
- command line query and export to file
- command line FTP

All of the above steps can be assembled into a single script file that can be scheduled or run with a single command. The ALA load process can load a single record if needed and can process updates as long as the record identifiers with the dataset are stable. Ongoing information loads from suppliers will be necessary to capture the addition of new information as well as changes (potentially corrections resulting from feedback by the ALA users) to information that has already been loaded.

When the ALA becomes aware that the information may be out-of-date, an update is requested from the provider (or the reverse situation where the supplier becomes aware that there are updates to be made to the ALA). Although no monitoring or scheduling infrastructure needs to be in place for this method to work, it requires manual oversight and in the long-term is prone to information becoming out of date. This is the only method possible until appropriate infrastructure is in place to automate the process. As changes happen in the source data, they are immediately applied to the target data stores. This type of update involves tightly integrated infrastructure and is ideal when information currency is important.

3.4.2.2.1 Scheduled updates

Updates from a data source are loaded into the ALA data stores on a regular basis. The frequency of these updates depends on the volatility of the data (how often it changes) and how the data are used (for example, information needed for a quarterly report does not need to be updated daily). Depending on which party is monitoring the currency of data, updates may be either sent from the supplier ('pushed') or requested by the ALA ('pulled'). An update 'pull' consists of the ALA requesting an update from the supplier. The supplier generates the update (either by responding to a remote query or running an export script) and uploads it to the ALA. If the ALA can send a trigger to the data store then this is preferable as the ALA will maintain the trigger and there is no responsibility on the supplier other

than 'listening' for the trigger. Due to IT security in most organisations, this scenario is less likely.

In a 'push' scenario the data supplier prepares and sends updates at regular intervals or when they choose to trigger them. When the ALA detects that a new dataset is available, the updates are integrated into the ALA. Initially the ALA will assist providers in setting up the scheduling of updates at their end and will set up systems in the ALA to notice updates and process them up—a 'push' scenario. The ALA systems check for updates includes an alerting system if an expected update is not found.

3.4.2.2.2 *Processing an update*

How an update is processed depends on the contents of the update. Updates are either a complete dataset or a delta (differential) load. A complete dataset contains a snapshot of the entire source database. A complete dataset is loaded into the ALA by deleting the entire dataset and replacing it with the new information. A delta load contains only those changes to the dataset that have occurred since the previous update. The delta load is processed as a set of insert new record, update existing record or delete record instructions that are processed against the existing data.

Delta loads are individually smaller datasets but are more complex to implement from both the provider and aggregator perspective. Delta loads are best used for larger datasets where a full delete and re-load would require too much time to process. Delta loads have the advantage of relatively easily allowing the tracking of the changes over time to a dataset. Changes can be extracted from a full delete and re-load, however, this requires a comparison of the previous and new datasets and the generation of a change log.

The decision to use delta or complete dataset updates will be made on a provider basis depending on the complexity and size of the dataset and the expertise and infrastructure available at the provider's end.

The following may be technical and resource constraints.

- Technical expertise: specific IT technical knowledge is required to set up the majority of the infrastructure to share data. Providers need to know how to install and configure a web server and software to serve data externally.
- Resource availability: providers that wish to share their data may not have the resources available to set up data sharing.
- IT policy and security: the IT security infrastructure and/or policy at the provider's site may restrict the software that can be installed and how information may travel into and out of their network.

The ALA has assisted with technical and resource problems using the following.

- Providing technical assistance: where a technically difficult sharing solution is needed, the ALA has provided technical expertise and resources to go on-site and set-up the infrastructure. This has been particularly effective with the organised museum and herbaria communities.
- Developing simplified sharing mechanisms: the ALA has developed simplified sharing mechanisms and processes that can be used without having to install and configure complex software. The infrastructure is able to transfer information without breaching the security policy of the provider. Although implementations are specific to the source systems, in general they involve the production of an export file that is sent via FTP to the ALA file server.

In addition to the legal concerns discussed previously, other potential data-related barriers include the following.

- Accuracy and currency: unless the data are current and accurate, it should not be used for decision-making.
- Value: the data may be used to derive income and the owners may wish to protect that income stream.

The ALA has assisted with data constraints using the following.

- Complete metadata: maintaining complete metadata on the authority of the resource and the accuracy and currency of the data allows people to make decisions on whether to use a particular dataset for decision-making.
- Self service: ultimately the process of sharing data should be fully automated, with new providers using an online tool kit to register, select their preferred terms of use and then map and upload their information into the aggregator.

3.4.3 Aggregation

Aggregation is the process of making the mobilised datasets available as a single dataset. For this to be possible the data need to be accessible in a single format (regardless of whether this is achieved virtually in a federated system or physically in a centralised system) and well-documented.

3.4.3.1 Metadata

The first stage in aggregating a dataset is to provide descriptive information (metadata) about it. This information provides context to the dataset including contacts and links to organisations and collections. Metadata may exist at multiple levels including for an institution, collection, dataset and individual records.

It may be a requirement of a particular analysis that data are selected from datasets with particular characteristics rather than based on record characteristics. This may include records from government departments which have undergone a vetting process or to distinguish survey or longitudinal datasets. Knowledge of the dataset as a whole may be needed to make an assessment of its usability.

Much of the search capability on records is related to the characteristics of the dataset including the owning institution, the project it was collected for the rights associated with the records, the

taxonomic coverage and technical details of the records. This information is often common to all records in a dataset and so is recorded as metadata rather than repeated for each individual record. A possible exception to this are datasets shared by aggregators; due to the variety of sources they have less features that are common to all records.

Where known, it is important to include the details of any information that is withheld (that is, exists in the source system but is not shared). There are several reasons why owners of data may choose not to share a subset of their records but including this in the metadata allows others to contact the owners for access. Data generalisations are also highly important for users—any reduction in the precision of the supplied records may affect their usability for analysis.

3.4.3.2 Standardisation

Standardisation is required for records from multiple heterogeneous datasets to be made accessible in a consistent manner that allows them to be queried via a single interface. Biodiversity information standards have been developed by international initiatives such as TDWG (<http://www.tdwg.org>). Standards deliver the advantages of providing data in a known format for data exchange and to facilitate automation. To be able to use a standard effectively, it needs to be understood. Even Darwin Core, possibly the simplest of the transfer standards, has over 150 fields. Effective use of standards such as Darwin Core requires selecting the appropriate fields to map to rather than trying to fill out the entire standard.

There are detailed schema specifications available for standards but limited documentation on how to use them. It may be considered ‘background knowledge’ in biodiversity data management but as the ALA develops re-usable systems to support more of a citizen science community, this type of documentation needs to be made available. Ideally, interfaces to data entry and mapping tools need to be designed with both a new and expert user in mind.

Standards also document controlled vocabularies. Without standard terms and scales it is not possible to use data from different sources in the same analysis. However, mapping terms to a common vocabulary is high value, but is an intensive task requiring domain specific knowledge.

3.4.3.3 Required fields

The minimum or required set of fields is the set where if there is no information in those fields, it is not possible to load the data. This threshold will depend on the purpose of the system and will be related to the most basic searching required. In selecting the minimum fields, an aggregator should keep the number and complexity of fields to a minimum although a balance between flexibility to integrate the maximum range of data and a basic level of quality is needed.

As an example, the required information to load a record into the Atlas of Living Australia is 'basis of record' and information on either a taxon or a location.

Basis of record describes what the record is of (for example, a specimen, occurrence, still image or sound). It is a quality-control field to differentiate between records with physical evidence that can be reviewed if needed and purely observation records. A missing basis of record from a dataset can often easily be resolved by setting a default value for the dataset.

Taxon, the organism the record is of, can be represented using a variety of scientific or vernacular terms including, scientific name, vernacular name and the higher taxonomy fields (such as genus, class and order).

Location can be defined using a variety of location fields including coordinates, Well Known Text, locality, State/Territory, references to geographic shapes and gazetteer entries. Some potentially useful information is not related directly to a specific taxon and so can be made available in relation to a

location. Some examples include habitat photos, literature, information on local naturalist groups and survey types.

The ALA uses the taxon concept—most basically as a scientific name, the namer and a date appropriate to the species name. Some cleaning may be required to make information suitable for integration with other sources and available for public release. As information may be mobilised from both primary sources and community hubs, records should be checked for duplication. Any processing carried out on a dataset should be recorded in the metadata.

3.4.3.4 Data mapping and transformation

To convert the raw data to a common model involves mapping the fields of the source data to the fields of the interchange schema. In some cases this will be a simple transfer, while in others it may involve complex data manipulation and processing. The more variety there is within fields, particularly for those that should have a standard format (for example, dates), the more difficult and time consuming the process will be.

Mapping and transformation requires not only technical skills to transform data but also knowledge of both the source data and target schema. The knowledge of both schemas is often not held by a single person as the person processing data at the aggregation point will know their standard schema well but have little knowledge of the source schema (the opposite is true for the source of the data). The impact of this is that mapping and transforming data is a collaborative effort between people at the source and aggregator involving back and forth discussion to establish the most appropriate fields to map to the standard schema and how to transform data without losing its real world meaning. To be able to carry out this role effectively, it is important for the staff working with data to have good communication skills, not only to maintain the relationships with the owners of the data, but also to ensure the data are mapped accurately.

Data lineage refers to the source of a particular value and what processing it has been through to reach the value it currently has from its original value at source. Also known as traceability, this is a vital capability of an aggregator that is intended for any kind of decision support associated with the data. Without traceability, users of the system cannot verify that original values have been mapped and transformed correctly. Users then may need to contact the custodians of the data to raise any concerns. Traceability is implemented in the ALA as an 'original' versus 'processed' view that shows the source values alongside the values as they appear in the ALA records. This function has been a key capability in resolving issues with ALA data.

3.4.3.5 Data ingestion tools

The most efficient tools for mapping data fields and transforming data primarily depend on the manner of accessing the data and size of the dataset. It may also be that the mapping is designed and prototyped in one system and implemented in another.

Applications (for example, Talend and Pentaho) designed specifically for mapping and transforming data can access a wide variety of data sources from text files to web services and databases, and map and transform the data to the same variety of outputs. There is an overhead in learning to use these tools and in setting up individual processing tasks but once created, the tasks are re-usable and often able to be scheduled. Extract Transform Load (ETLS) tools are efficient for complex, large volume, repeated data loads where the initial outlay in time to learn the tool and create the processes is offset by repeated use.

MS Access is a common platform for ad hoc (and even some non-ad hoc) databases and may be used as a transfer platform (that is, a dataset may be sent as a .mdb file). MS Access can also be a useful utility for connecting to some file types (for example, .dbf files that hold the data for shape files) and reviewing text or spreadsheet files that are too large to be

viewed in older spreadsheet programs. Export capabilities are also configurable. There is an open source alternative called Base in the Open Office package.

Spreadsheets are commonly used for small projects and relatively simple data manipulation. Their key feature is the ability to interact with data directly rather than via queries. Small datasets received in this format are easily manipulated within the spreadsheet, although it is important to retain an original copy for traceability. There is an open source alternative called Calc in the Open Office package.

Data may be transferred as a SQL dump and require a database to load it for processing. If no ETL tool is available and there is expertise in the particular database then it is possible to load in data from other sources and transform them using queries which are reusable and can be retained for traceability. Command line scripting of database functionality is useful for automation. This is a fairly heavy-handed way of handling smaller datasets and should be reserved for uses similar to ETL tools. Open source alternatives include MySQL and PostgreSQL.

Whether a system specifically designed to harvest a particular web service (such as TAPIR, DIGIR or BioCASE) or a custom-coded application, these systems are more efficient for maintaining the currency of data via automated updates than initial loads or full refreshes of data. Access to web services typically returns highly verbose data (usually XML) in small record sets several hundred to a few thousand at a time. Transferring a dataset of several hundred thousand records or more can take days to request, download and process. Gaining permission to access the data via the access point can also be problematic from an IT security perspective.

3.4.3.6 Discoverable data

Once an aggregated dataset has been assembled, users need to be able to find these records. This functionality is provided by creating searching and

filtering capability based on indexes across fields of interest. Initial indexes are provided on the required fields and other indexes are created based on the available data and user interest. Creating the indexes will expose gaps in source data, or at least the data that was provided, and may prompt more complete data in future updates.

Data quality is a primary consideration of users when they are selecting which data to use for their analysis. Further principles and definitions are outlined in the appendix. Data integrity considerations consist of whether the record is cohesive in terms of the field content and whether the information makes sense or is usable in a real world context. This can be considered at any of the steps in the life cycle of a record from the original source to production of an export, import into another system or downstream processing.

A record with good integrity will have data in all appropriate fields and the data will conform to best current practice standards. Data values should be within specified bounds. Unless data meets basic integrity criteria, it should not be loaded and needs to be referred back to the source. In the ALA case, data will be loaded as long as it at least has either a recognisable taxon it can be linked to or a usable location.

How information will be used determines what constitutes a measure of quality in a particular context. To service the widest range of applications, users should be able to evaluate the fitness-for-use, or 'usability' of data. It is the users, rather than the aggregator, who need to determine which data will meet their quality threshold for a particular use. It may be possible to improve the quality of some data parameters from the context of other parameters. For example, the accuracy of a location may be improved using locality descriptions and known ranges for a species.

The usability of data is based on metadata as well as core measures such as the accuracy of the geospatial coordinates versus the coordinates themselves. It is important to ensure that the metadata is available and able to be used as a filter when selecting data for inclusion in an analysis.

Wherever possible, metadata should be collected when systems and records are created to ensure it is as accurate and as complete as possible. For older systems and records, metadata may need to be generated using the best available contextual information. It is vital that all derived metadata is flagged as such and the methods used to produce it are exposed.

3.4.3.6.1 Persistent identifiers

The use of persistent and preferably globally unique identifiers prevents duplication across systems and allows different versions of a record to reference their source. If a source system changes its database platform or owner, the original records are still able to be referenced.

3.4.3.6.2 Licensing and attribution management

The terms of use associated with a dataset or record should include usage (if not necessarily quality) considerations. The license associated with a dataset needs to be available as a filter term.

What makes up a good record? For records that will be used in analysis rather than as a description, there is a value for each of a set of core measures (without which there is effectively no record) and extensions providing context to the core set. To assess usability, each of the core values needs to have metadata including:

- an indication of the accuracy and precision
- information on verification or evidence for the value and accuracy (how can I check the value or have a confidence in it?).

A good record allows fitness-for-use to be determined. The presence or absence of values in any of these groups, as well as the values themselves, should be available as filters on data. It is important to be able to determine the difference between 'value not displayed' (possibly for sensitivity reasons), 'no value available' and '0' for example.

The key unit of analysis in the ALA is the occurrence record. The core values of an occurrence record include the 'what', 'where' and 'when'.

What

- value: species
- precision: to what taxonomic rank has the identification been made
- accuracy: a term indicating how much confidence there is in the identification
- verification: basis of record (one of: observation, photo, specimen, sound, or footprint, among others)
- identification details (who, when, references).

Where

- value: coordinates, locality and location description
- precision: how precise is the way of measuring location (for example, GPS coordinates reported to five decimal places)
- accuracy: margin of error in the location (within a 100m square from the point)
- verification: GPS, map, accuracy reduced for sensitivity reasons, others.

When

- value: date, time
- precision: to what level was the time recorded (time, day, month, year and decade)
- accuracy: over what time period could the event have occurred (one hour, one day etc.)
- verification: survey trip dates, diary/notebook entry, or trap placement period.

The fields to record this information are for the most part available in Darwin Core. Any that are not available (for example, date accuracy) will be raised for consideration as an addition to the ALA data model and Darwin Core standard itself (see <http://www.tdwg.org>). This grouping of fields into values, precision, accuracy and verification types may be used as the context for documentation explaining how to use Darwin Core fields. This grouping can also be used in the interface design for data entry and mapping tools. For example, when recording a sighting, the researcher should indicate the certainty of the identification (selecting from options), the basis for the identification (selecting from options) and how the identification was determined (selecting from options).

Data entry systems developed by the ALA encourage (but not necessarily make mandatory as the information may not be available) comprehensive resource metadata for datasets. Data entry tools should encourage the entry of usability metadata at record level or through the setting of defaults that apply to sets or collections of records such as surveys, time periods or users.

3.4.3.6.3 Legacy data

To facilitate quality records, existing data requires:

- digitisation (if not yet digitised)
 - entry of quality data and metadata when digitising records
- validation
 - identification of error types and recognition methods
 - review of existing data and metadata against quality model and measures
- cleaning
 - contact data sources to complete records
 - derivation of data and metadata unavailable from source
- mobilisation either before or after validation and cleaning.

Validation (quality review) of legacy data aims to bring previously collected data up to current quality standards. This will be an ongoing task that ALA systems need to facilitate while making the data available for use. Validation tools focus on identification of gaps in quality so they may be able to be addressed in the future. The second part of the process is filling or repairing gaps and errors. Missing data should first be completed by identification and updated from the original source data. If the data are not available from the source, it may be able to be updated by an expert or through the application of algorithms and standard values. If no other values are available, fields should be flagged to indicate that efforts have been made but no value was found. This hopefully reduces the same issue arising many times. Validation and update processes may be carried out at the source or the ALA. If validation occurs on ALA infrastructure then the data are available and may be suitable for some applications even if it is not complete.

The completion of records with source data or update by an expert is a resource and expertise intensive task and will be ongoing. Research and collecting institutions are consistently resource poor and the review and update of data is usually a low priority. The data are collected for a purpose and once that purpose is fulfilled, interest in the data is usually of low priority. As a result, the validation and record update systems developed by the Atlas focus on identification and flagging of gaps and potential errors but do not rely on the issues being immediately addressed at the source. The traditional goal of a closed feedback loop may not prove to be practical. ALA systems should aim to display multiple versions of records so that corrections are available even if they have not yet been updated at the source.

3.4.3.7 Sensitive data

Some datasets, records or fields within records may be considered sensitive—that is the release of the information may cause harm. The impact of releasing

the information may relate to species conservation, bio-security or personal privacy. The determination of sensitivity is maintained according to governmental legislation. There are other reasons information may be withheld including contractual obligations, research or commercial advantage but these are decisions made by the owner of the data rather than legislative requirements.

It is often the case that legislated sensitive data are processed according to rules provided by the relevant authority after it is shared. The primary motivation for the checking and processing of records is for the case where the creator of the record is not aware of the sensitivity of their information. For example, a naturalist recording species observations from New South Wales may not be aware of the sensitivity status of a particular species in Western Australia. As with many components of biodiversity information, the legislation defining sensitivity may have temporal and geographic boundaries. Some examples include:

- a particular state after a particular year when the species was determined to be a conservation risk
- any time except during a short period when there was an outbreak that was successfully contained
- a pest exclusion zone
- following the introduction or update of legislation (for example, privacy or copyright).

3.4.3.7.1 Privacy

Federal, State and Territory privacy (or similar) legislation provides direction on the release of an individual's personal information. As this issue was being raised broadly, the ALA approached the privacy contact indicated for each State with a general enquiry as to the privacy legislation in their jurisdiction (specifically relating to the museums and herbaria). Excerpts from their responses include:

From the office of the New South Wales Privacy Commissioner:

'...if the information is about a person deceased for more than 30 years it will not be within the definition of "personal information." The PPIP Act will thus not apply.'

From the Office of the Information Commissioner (Queensland):

'...Nor does privacy legislation apply to deceased persons...'

The Office of the Queensland Privacy Commissioner also went on to offer the following practical advice:

'In terms of practicality, this is an area where if the personal information of an individual is 'published', the individual would rarely object to the publication and if they did, the 'damage' suffered by the publication would be likely minimal. It would be impractical if not impossible for your agency to contact everyone concerned to obtain consent. Accordingly, we suggest two courses for the future:

- *Provide a collection notice on all forms used from this point forward including an opt-in capacity for the information that is transferred outside Australia.*
- *Include a prominent feature on your web-site publicising that historical personal information is included and providing the mechanism to individuals who object to their inclusion or who wish to contest the accuracy of the information to have the information removed or corrected.'*

As of this publication date responses have either not yet been received from the offices in other jurisdictions or they indicated their office could not offer any advice to the ALA. The advice offered by the Queensland Privacy Commissioner presents a practical approach to this issue that the ALA hopes will be adopted by other data providers following their own enquiries as to their individual

circumstances and the operation of any privacy legislation in their jurisdiction. There appears to still be significant investigation needed by individual institutions in each jurisdiction. The collection of institutions in a particular jurisdiction may wish to consider making a general enquiry to their information or privacy commissioner.

- <http://www.privacy.gov.au/>
- <http://www.privacy.gov.au/law/states>

An internal review and update of institution collection and privacy policies to request permission to disclose collector names (the use of an appropriate 'collection notice') will avoid this issue continuing into the future. If this issue cannot be resolved to allow the full records to be shared, the simplest solution is to not share the portion of the records that contain personal information. The result of this is reduced quality of the data available.

3.4.3.7.2 Processing sensitive data

Typically processing of sensitive data involves reducing the precision of some values in the records or in extreme cases withholding the records entirely. Data can be de-sensitised by lowering its accuracy to a nominated resolution either geographically or temporally. The co-ordinates may be made less accurate or the ALA may have data that are several months old. Rules make it possible for the ALA to publish sensitive data and also allow the owners to maintain access control of the most valuable data. Whatever the degree of transformation applied to the records, it is vital to clearly indicate that the records have been processed.

Examples of sensitive species lists can be found at: <http://collections.ala.org.au/datasets#filters=contentType%3Asensitive+species+lists>

3.5 Recommendations

The areas outlined below would have a substantial impact on the availability of biodiversity data for sharing.

3.5.1 Supporting systems

As distinct from digitising and georeferencing legacy data, the capture of 'high quality' data should be facilitated as it is collected. Developing a highly extensible and configurable data collection platform requires the collection of the key elements of biodiversity data and all the useful supporting information around those key elements 'behind the scenes' e.g. the georeference protocol and precision and uncertainty of locations. Extensibility allows the collection of as much additional information as needed for specific projects.

The key to successfully establishing quality data collection is to minimise overheads. System interface and workflow design is vital to the success of the tools. The entry of quality data at the point of collection will minimise validation and cleaning required. Reuse of repeated information through the creation of templates helps to facilitate this.

Systems that facilitate the reuse of methodologies and templates where possible allow for greater interoperability of datasets. For example, a reusable template for a particular collecting methodology that includes all the repeated metadata may be:

- methodology
 - collection methodology
 - basis of record
 - collector name
- taxonomy
 - species pick lists based on area checklists or taxonomic focus

- identification
 - person who identified it
 - references
- location
 - coordinate accuracy and precision
 - how the location was measured (for example, GPS make and model, map name and scale).

The information above may be systematically applied to records and updated if there is a difference with a specific record. The entire template can also be reused for future events.

3.5.2 Cultural change

An important principle is to foster a culture of availability and free and open access to biodiversity data among the groups that collect it (including research, education, government, citizen scientists, management and policy). Ideally, there should be policy and/or legislative backing for the sharing of data and provision to a national data aggregation point. Awareness of the requirements for interoperable data (for example, full records with supporting metadata and traceability) should be increased with those that collect, manage and digitise data. This encourages the consideration of sharing data when designing experiments and projects including appropriate legal notices regarding sharing data and the release of personal information (the collector name associated with the records) at the time a project is planned. One potential method of facilitating cultural change is to incorporate principles of open access and interoperability into education including graduate and post-graduate research and incorporating these principles into funding requirements.

3.5.3 Manage data as data

Although occurrence and ecological data (unit facts and figures) are now usually captured in a system of some type (primarily spreadsheets) information surrounding the data is usually captured and managed as a document. Unfortunately this makes it very difficult to extract and use this information for further analysis. The primary example of this is species descriptive information (such as the morphological or behavioural characteristics of a particular taxon) that is documented in a paragraph of text. If the characteristics were available as units of data then they could be used to provide additional analytical capabilities around the occurrence data.

4. Atlas of Living Australia data and tools

4.1 Introduction

The Atlas of Living Australia (ALA) aggregates an extremely broad range of biological and related data from an equally wide range of data providers. The core biological data in the Atlas are species occurrence records from human observations and collections in the Australian region. These records cover all taxonomic groups, terrestrial, marine and freshwater habitats and from the year 1629 to a few minutes ago. These data have come from herbaria, museums, citizen science groups, special interest groups, private individuals, State and Territory bio-surveys and university departments (see <http://dashboard.ala.org.au>).

Biologically-related data in the Atlas include:

- environmental and contextual data layers (from a wide range of agencies)
- images and video (via Flickr and Morphbank)
- biodiversity-related literature (Australian Biodiversity Heritage Library)
- DNA data (Australian Barcode of Life)
- identification keys (Identify Life)
- species interactions (recently initiated)
- gazetteer (Geoscience Australia, Global Administrative Boundaries, contextual layers).

4.2 ALA users

The community served by the Atlas ranges from primary school students to research scientists. A survey of registered users suggests that half the ALA's audience are professionals in research, consulting, land management or policy development. The remainder are either amateur naturalists or those interested in the environment from an educational perspective. In the Atlas, the species pages for example cater to general interest while predictive modelling is aimed at land managers and research needs. For example, a paper by Booth et al. (2012) shows how the ALA can be used to select species for planting as part of the \$946 million

Biodiversity Fund. More than 50 natural resource management agencies have requested information on how to use the ALA for this purpose (<http://www.ala.org.au/blogs-news/data/using-the-ala-to-help-develop-biodiverse-plantings-suitable-for-changing-climatic-conditions/>, last accessed 1 May 2013).

4.3 ALA data types

As with the proceeding discussions, the following discussion regarding the biological data in the Atlas of Living Australia is a snapshot in time.

4.3.1 Taxonomic groups

Records are being added to the Atlas every day. Species records in the Atlas cover most taxonomic groups. Figure 1 shows the number of records at the kingdoms level. The distribution is far from equal, with Animalia and Plantae covering ~93 per cent of the 33 million ALA records. At the other extreme, there are currently only seven records of viruses. In some cases, it is anticipated that specialist collections will help to improve the balance. For example, an agreement has been signed with FungiMap (<http://www.rbg.vic.gov.au/fungimap>) to include all the records from their database into the Atlas.

4.3.2 DNA sequences

It is anticipated that the proportion of DNA records in the Atlas will steadily increase over time. At the time of writing, records in the Atlas from the Australian Barcode of Life Network are summarised in Figure 2.

4.3.3 Life forms

The Atlas also provides a breakdown of the records on the basis of life-forms. This adds an alternate perspective on taxonomic coverage (Figure 3). Note that there is no universally accepted vocabulary or ontology of life-forms. The grouping selected for use in the Atlas seemed convenient for public understanding as such a breakdown would be of lesser value to the scientific community.

4.3.4 Names

The Australian National Species Lists (NSL: <http://biodiversity.org.au/confluence/display/bdv/NSL%2bServices>) is a joint project between a number of Australian museums and herbaria, universities, CSIRO and the Australian government. The National Species Lists will compile all existing names for Australian biota (including linking of names that are synonyms) as an essential support for ALA tools for mapping and modelling species distributions.

Figure 4 displays summary statistics of the NSL at the time of writing. The Royal Botanic Gardens (RBG) Melbourne has been contracted by the Council of Heads of Australasian Herbaria (CHAH) to administer the Australian National Species Lists for plant names. The NSL not only integrates a wide range of taxonomic names, like most functions in the Atlas, the NSL also provides web services to other websites to enable a comprehensive range of search and display functions relating to species names.

4.3.5 Basis of record

'Basis of Record' is a TDWG Darwin Core standard field (see <http://rs.tdwg.org/dwc/> and more specifically <http://code.google.com/p/darwincore/wiki/RecordLevelTerms>) that provides a description of the nature of the biological record. As Figure 5

shows, most of the records in the Atlas are either 'Human observations' (~70 per cent) or 'Preserved specimens' (~20 per cent). Figure 5 does however show that the Atlas encompasses a broad range of record types and therefore demonstrates the complexity of the infrastructure that is the Atlas of Living Australia.

4.3.6 Taxonomic keys

The Atlas of Living Australia supported the establishment of the international project <http://www.identifylife.org>. There are currently 532 keys in Identify Life (Figure 6). These keys or traits are used for the identification of species or high taxonomic groups such as centipedes of Australia.

4.4 Spatial coverage

As noted elsewhere, the ALA has a focus on the 'Australian region', but not exclusively so. For example, the inclusion of invasive species in their home range into the Atlas enables spatial modelling applications (see Figure 7) noting that transparency has been lowered to better identify the continents). The definition of the 'Australian Region' is also loose in that it notionally covers mainland Australia and its islands, External Territories (such as Australian Antarctic Territory), Coastal Waters, the Continuous

Figure 1. Atlas records by Kingdom.

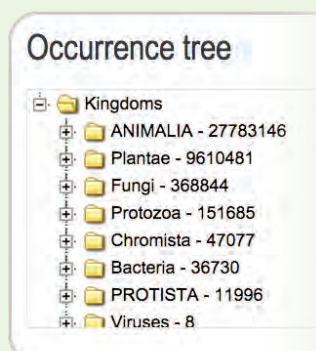


Figure 2. DNA records in the Atlas.

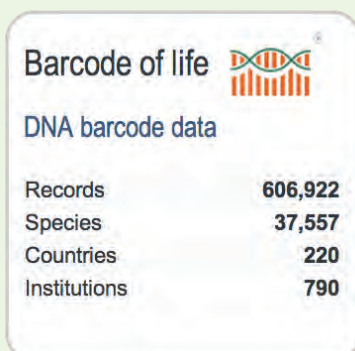
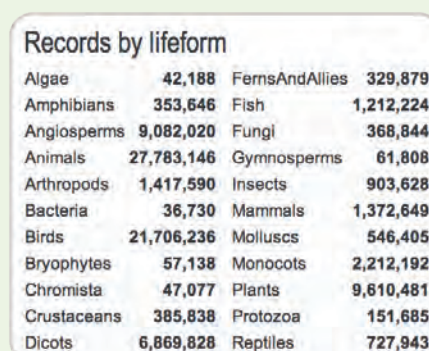


Figure 3. Atlas records by life-form.



Zone, Continental Shelf, Territorial Seas, Exclusive Economic Zones, Australian Fishing Zone and oceanic areas. This region is roughly nine per cent of the surface area of the planet.

As can be seen in Figure 7, species records encompass terrestrial, marine and freshwater environments. While the Atlas can discriminate between terrestrial and marine habitats by contextual and environmental layers (see below) or taxonomy, delineating the freshwater environments is currently not possible. The Atlas has received a list of species that have at least a freshwater phase but this list is not comprehensive nor is such a list managed by the limnetic community. What this means is that freshwater species occurrences may not be easily recognized by non-experts, and currently cannot be associated with freshwater environmental data.

The ALA may have a useful role in setting up the coordinating data management infrastructure for science groups to interact and provide authoritative advice on classifications of species that are of national and scientific importance.

It is a simple process to download international species occurrences from GBIF (<http://www.gbif.org>) or from your own data and import them into the Atlas. Once these records are imported, they

are treated the same as any other Atlas occurrence. This permits amalgamation of external data with ALA records. Imported records are available for a session and are not integrated with the ALA database.

Figure 8 shows a snapshot in time of the occurrence density of biological records in the ALA. Note that the distribution shows:

- centres of species richness
- the density of terrestrial records tends to reflect the populated areas and the road network
- marine hotspots around Heard and Macquarie Islands
- Antarctic 'hotspots' adjacent to Australia's Antarctic stations
- Southern Ocean 'hotspots' that reflect regular Antarctic ship tracks.

While one would anticipate that the records in the ALA would reflect the area of the States and Territories, records better reflect the population centres within the States and Territories. In outback areas, occurrence records roughly delineate roads and tracks.

Table 2 shows the number of record occurrences and species richness by State and Territory (sorted by area). Note that the correlation between species

Figure 4. Atlas records by species names.

National Species Lists	
Accepted names	195,424
Synonyms	146,974
Species names	152,080
Species with records	111,566

Figure 5. Atlas records by basis of records.

Basis of records	
Human observation	29.14M
Preserved specimen	8.33M
Machine observation	321,614
Genomic DNA	154,529
Living specimen	59,890
Fossil specimen	22,114
Image	19,939
Nomenclatural checklist	5,923
Sound	4,557
Literature	575
less..	

Figure 6. Identification keys in the Atlas.

Identify Life	
Identification keys:	
	532

Figure 7. Spatial coverage of Atlas records.

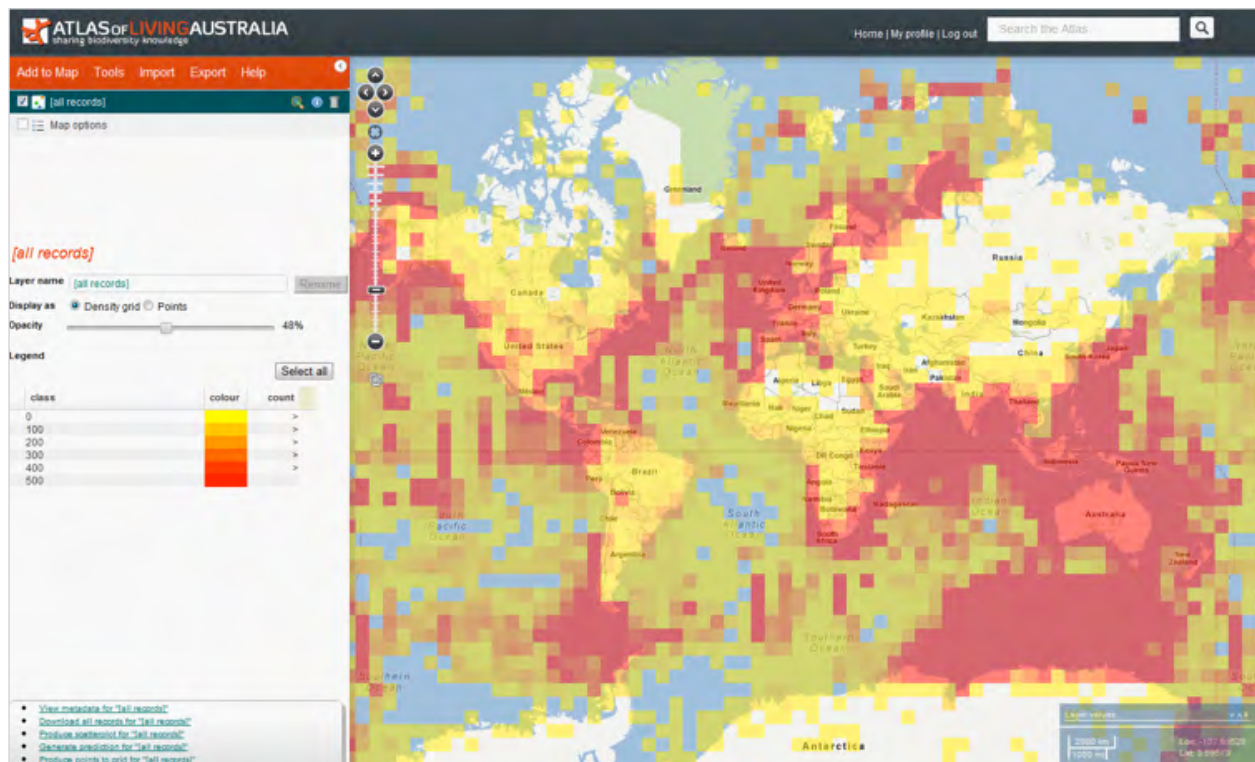
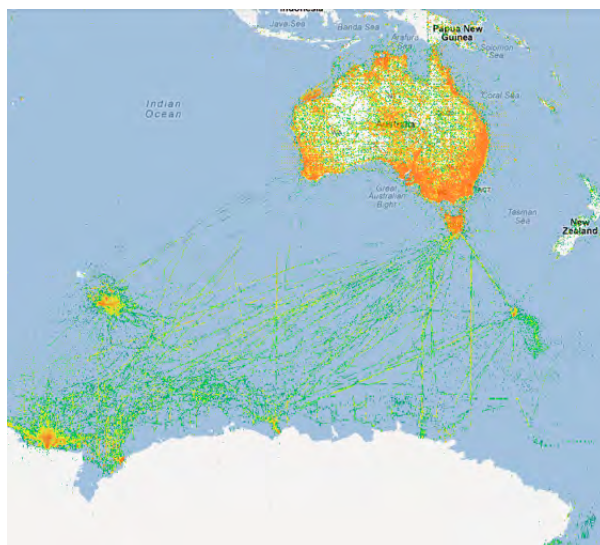


Figure 8. Occurrence density in Australian region.



occurrence records and species richness is high ($r=0.85$) as may be anticipated; the more records there are, the more species there are in those records. There is no significant relationship between species occurrences and area (or richness and area). Table 2 reflects what States and Territories have currently supplied to the Atlas, not what bio-data they may have in their own records.

Table 2 covers the terrestrial and aspects of the freshwater environment. The ALA has approximately 10 million records covering the marine environment, external and foreign territories. Given that species generally have no respect for political boundaries, a better idea of spatial coverage may be gained by examining their distribution by bioregion.

Table 2. Records by States and Territories.

States/Territories	Total occurrences	Total species	Total area (sqkm)	Species/Area
Australian Capital Territory	1,460,202	6,936	2,419	2.87
Tasmania	582,071	20,615	67,324	0.31
Victoria	3,705,579	49,003	227,343	0.22
New South Wales	6,446,950	80,901	802,588	0.10
South Australia	3,568,908	31,584	985,408	0.03
Northern Territory	1,708,691	19,381	1,351,902	0.01
Queensland	3,098,524	64,685	1,736,618	0.04
Western Australia	2,346,444	42,219	2,532,655	0.02
Total	22,917,369	315,324	7,706,258	
Minimum	582,071	6,936	2,419	0.01
Maximum	6,446,950	80,901	2,532,655	2.87
Mean	2,864,671	39,416	963,282	0.45
Standard deviation	1,691,084	23,266	829,602	0.92
r(Occurences-species)	0.85			
r(Occurences-area)	0.10			
r(Species-area)	0.33			

4.5 Temporal coverage

The earliest record in the ALA is 4 June 1629 and the most recent will usually be within the last 24 hours. Figure 10 (in Section 5 Business requirements elicitation) and Table 2 show all Atlas bio-records and the number of species (species richness) over time. Note that the scale on the graph is log(10) to enable a clearer idea of the smaller number of occurrence records in earlier periods. Note that the temporal correlation in Table 3 of species richness between all species and *Eucalyptus* species is a high 0.98 (see also Figure 9).

4.6 Methods for data entry

Biological records can be added to the Atlas of Living Australia by the following mechanisms.

- Collections of themed data submitted from agencies such as herbaria, museums and similar agencies, citizen science or specialist groups and individuals. These collections could be in a wide variety of formats, from text and spreadsheet data to Darwin Core records generated from the Biological Data Recording System (BDRS)—an open-source (and freely available) toolset suitable for a wide range of field data capture of biological data.
- Occurrence records from State and Territory government agencies. These will include *ad hoc* observations and systematic surveys.
- Checklists of species within a defined set of polygons from State and Territory government agencies, citizen science or specialist groups and individuals. For example, the ALA has checklists of coral species within a set of world coral eco-regions.

Table 3. Temporal coverage of records and species for all species and *Eucalyptus*.

	All species		<i>Eucalyptus</i>	
	Records	Species	Records	Species
Series	1	2	3	4
<1850	36552	9950	898	203
1850s	15638	6217	282	109
1860s	19966	7391	209	92
1870s	29871	10075	382	151
1880s	77657	15395	1063	252
1890s	105083	19177	2153	295
1900s	152004	23950	5424	444
1910s	139872	22246	4612	435
1920s	123283	23741	3531	498
1930s	177303	25998	4533	433
1940s	185282	27503	5238	529
1950s	497013	35329	9598	645
1960s	1047644	48038	22196	805
1970s	2544757	58640	40269	878
1980s	3926474	61652	52050	951
1990s	5558689	57937	57470	966
2000s	10494967	47665	44611	947
2010s	1162018	14319	1032	291
r(Records)	0.816751			
r(Species)	0.98025			

- Polygons defining the spatial distribution of fish species by marine biologists. These are a special case of the species checklist with the difference being that the 'expert distributions' relate to a single species within a defined area.
- *Ad hoc* occurrence observations entered directly into the Atlas via the web or by Atlas software on mobile devices using iOS or Android apps such as OzAtlas.

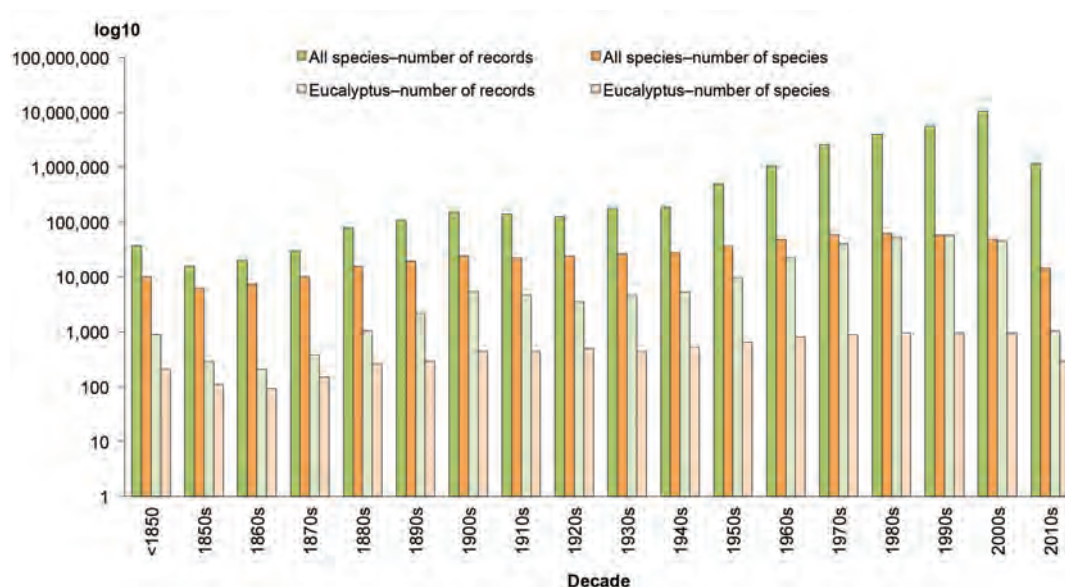
4.7 Fitness-for-use/data quality

While the term 'data quality' has been widely used, it is far preferable to use the term 'fitness-for-use'. Data that may be considered of 'poor quality' for some application may be more than suitable for others. The general principle behind the Atlas generally, and more specifically the spatial portal, is to expose the data as it was received and to flag any issues that can be detected by various rules and algorithms. For example, the record of a terrestrial species occurring in the ocean will be flagged but the coordinates will not be altered.

All data are checked for general validity and subjected to a suite of automated tests implemented by Bennett and Belbin. For example, a suite of five reasonably independent terrestrial climate variables are used to detect environmental outliers. Taxonomic checks are also performed against the National Species List. As noted above, questionable data are flagged rather than omitted. It is up to the user to determine the fitness-for-use of each record used in the Atlas (see <http://www.ala.org.au/about-the-atlas/how-we-integrate-data/data-quality-assurance/>).

Work is continuing to improve the detection of data issues, for example, the identification of 'duplicate' records. Records may come direct from the person, organisation or project that collected the data and via an intermediary. For example, a record of a marine species may come directly from the University of Tasmania as the collector and also via OBIS (the Ocean Biogeographic Information System). In some cases, value may be added to data so automatic deletion of duplicates is not advised. When data are used in any way from the Atlas, it should be downloaded and the attributes of all records carefully examined to detect potential issues that may affect an analysis.

Figure 9. Temporal coverage of records and species for all species and *Eucalyptus*.



4.8 Sensitive data

A comprehensive Sensitive Data Service has been developed for the Atlas. This service entailed integration of the status of all sensitive species by each State and Territory. A set of rules have been generated that control how sensitive data are handled and displayed within the Atlas. In some cases, for example, records will have their coordinates generalised to indicate the region but not the exact location. In other cases such as spatial prediction, sensitive data will be automatically omitted and the user advised.

At the time of writing, sensitive data can take two forms within the Atlas:

- Records that have been processed by the data provider or the Atlas and flagged accordingly as 'Information Withheld' or 'Data Generalised'
- Records that have been changed by the data provider and have not been flagged as such by the Atlas.

Case 1 is where rules have been applied in accordance with the data provider (held in the

Sensitive Data Service) to produce public records via the Atlas. In this case, the Atlas flags the processed records with values in two additional record fields: 'Information withheld during processing' and 'Data Generalised during processing'. An example of the former would be 'The event ID and day information has been withheld in accordance with Birds Australia data policy' while the latter may be, for example, 'Coordinates generalised to 0.1 degree in accordance with the State of Victoria policy'.

Case 2 is where data provided to the Atlas have already had coordinates generalised in some way. For example, latitudes and longitudes of some of the records of *Leipoa ocellata* have been rounded to one decimal degree. In this case, as the Atlas has not performed any processing of the records, no flags are currently set. This issue has been raised and the intention is to see if such rounding can be detected and to provide a suitable flag.

At the time of writing, 204,856 records were (spatially) 'generalised' by the Atlas and a further 56,827 records were already generalised when received by the Atlas. As noted above, this does not

currently take account of records that appear to have been generalised but not flagged as such. Limited information is known about this latter category. It is likely that around one per cent of the records in the Atlas have had their coordinates generalised in some way.

4.9 Environmental data

Most of the environmental data in the Atlas are terrestrial (292 layers) but recent efforts have added 22 marine layers (Appendix 3). However, no systematic environmental data currently exists for the freshwater environment. Each State and Territory monitors and maps different environmental parameters. The AUSRIVAS initiative (<http://ausrivas.canberra.edu.au/>) has been moving toward providing a standardised approach. The emerging Australian National Aquatic Ecosystem Classification Scheme (currently version 0.6) also looks like becoming an effective way of classifying freshwater and related environments.

Early in the project, two workshops were conducted to determine what tools should be available in the Atlas and what associated environmental data would be required (see Flemons & Belbin 2010, Flemons et al. 2010). Subsequently, it was realised that the Atlas would never have the resources to be able to implement tools that satisfied all the requirements of the research community and secondly, that community usually wanted to apply their own desktop tools anyway. It was therefore concluded that the ALA should provide a suite of tools that demonstrated the added value of bringing disparate but related biological and environmental data together. In addition, the emphasis was moved to more comprehensive support for *ad hoc* data import and export.

While the ALA is the national repository for biological data, environmental data may be better held by agencies such as Geoscience Australia or the Bureau of Meteorology who routinely handle such data.

However, the problem was that no other agency was in a position at the start of the Atlas project to gather, integrate and deliver these data to the Atlas. This remains the case even now. To enable the Atlas to demonstrate the utility of integrated biological and environmental data, it was therefore left no option but to establish an Environmental Data Library.

The library has been established and managed by the Geospatial team. Ideally, a small cross-institution committee should manage this library. Regular reviews are required of potential new layers and the utility of existing layers. Useful new layers (e.g. Williams, Belbin et al. 2012), updates to existing layers and deprecation of layers are inevitable as the technology improves.

5. Business requirements elicitation

5.1 Introduction

The requirements for biodiversity environmental information that contribute to policy-focused decision-making are diverse. As a component of this project, the project developed and tested a semi-structured approach for eliciting stakeholder requirements around biodiversity information. The purpose of the exercise, in the form of a workshop, was to determine whether it is possible to apply an approach for rapidly eliciting decision-making requirements from stakeholders, framed around what is possible to deliver (feasibility).

The feasibility assessment was the responsibility of ALA and CSIRO while stakeholders from the Department of Sustainability, Environment, Water, Population and Communities provided a policy and decision-making perspective to the workshop.

The primary objective of the process was to design, implement and test a method for eliciting business needs rather than transitioning the high-priority needs to products and services. The workshop and views expressed were framed around this 'experimental' construct. This was important as the policy 'actors' in the workshop were a small cohort from a relatively large agency with diverse needs for biodiversity information. Similarly, the 'provider' actors represented only one capability (ALA and CSIRO) at the expense of a more comprehensive treatment of possible datasets and modelling approaches.

The following section describes the method developed to elicit preferences, presenting some of the experimental results and reflections on the exercise and providing recommendations for future priority setting. The later is particularly relevant to activities around priority setting being led by the Bureau of Meteorology as one of its contributions to the National Plan for Environmental Information Initiative.

5.2 Workshop objectives

Using a participatory approach, the workshop sought to:

- engender a shared understanding of policy requirements and their relative importance.
- reach consensus on the highest priority and most feasible requirements informed by a simple conceptual framework.
- define parameters around the high utility and high feasibility products (1–2 products).

5.3 Detailed method

Business requirements involved an initial scan (Stage 0) conducted ahead of the workshop, and four key stages conducted during the workshop.

- **Stage 0 (Initial scan):** Two policy actors were provided a template to provide an initial scan of needs around biodiversity information ahead of the workshop. This ensured policy actors came prepared to the workshop and allowed them to consult internally around needs prior to the workshop. Both responses were integrated into an individual spreadsheet for people in the workshop (see [Table 4](#)).
- **Stage 1 (Policy ranking and provider feasibility):** Policy actors were provided 15 minutes to rank their requirements and providers conducted an initial feasibility evaluation (Yes/No options). Provider feasibility was evaluated by rapidly assessing whether data was available in ALA to support policy needs and whether analytical and modelling methods were sufficiently mature to analyse the data.
- **Stage 2:** A conceptual model framing biodiversity data against possible decision-making needs was presented to stakeholders to inform their thinking in the forthcoming sessions. The conceptual model is presented in Section 6 ([Figure 11](#)).

Table 4. Examples of initial-scan business requirements shown against policy/program drivers. The examples represent a small cross-section of a more comprehensive list used in the workshop.

Example business requirement	Policy/Program/Driver/Outcome
Understanding and reporting on change in the extent and distribution of native vegetation and changes in the extent, frequency and duration of inundation of aquatic ecosystems.	Australia's Biodiversity Conservation Strategy 2010–2030, Native Vegetation Framework and associated goals and targets; National Wildlife Corridors Plan; Caring for Our Country and Biodiversity Fund.
Understanding and reporting on changes in the condition of native vegetation and aquatic ecosystems.	Australia's Biodiversity Conservation Strategy 2010–2030, Native Vegetation Framework and associated goals and targets; National Wildlife Corridors Plan; Caring for Our Country and Biodiversity Fund.
Understanding and reporting on changes in the fragmentation and/or connectivity of native vegetation communities and aquatic ecosystems.	National Wildlife Corridors Plan; Australia's Biodiversity Conservation Strategy 2010–2030, Native Vegetation Framework and associated goals and targets; Caring for Our Country and Biodiversity Fund.

- **Stage 3:** For the top four policy requirements which were regarded as feasible, each priority was explored via a 15 minute question and answer session.
- **Stage 4:** A group discussion explored the two highest scoring policy needs in further detail to develop high-level specifications describing some simple parameters such as spatial extent, temporal repeat and accuracy.

5.4 Discussion and recommendations

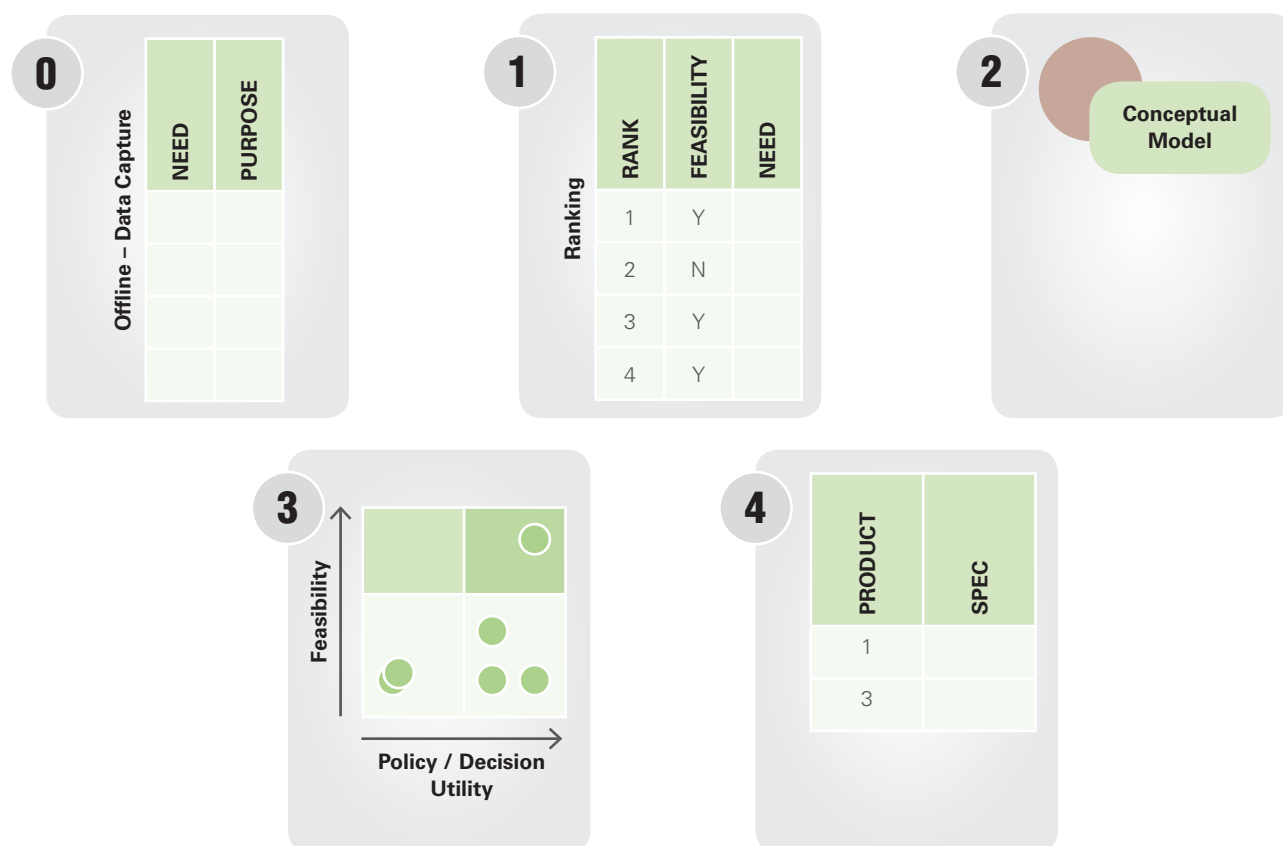
Relying primarily on a workshop setting to elicit preferences to evaluate feasibility and define some high-level parameters around priority products and services remains challenging. In this instance the three hour format of the workshop placed added pressure on realistically being able to realise all objectives. In particular there was limited opportunity to provide a detailed treatment of Stage 4 (identifying more detailed parameters around priority products and services). Second, 'scoring' feasibility was also challenging as this required a more structured approach rather than relying on subjective assessments of feasibility.

This business requirements experiment highlighted that workshop objectives were ambitious. A more achievable objective was to treat the workshop as a forum to gain an improved understanding of stakeholder perspectives and key issues, rather than necessarily attempting to identify and detail the one or two priority biodiversity products and services. Nevertheless, framing the workshop around this later objective was sensible as it encouraged the workshop to traverse the key issues for both providers and users. The following describes some of the successful elements of the process, some of the challenges, and offers recommendations for future priority-setting processes.

5.4.1 What went well

- Conducting an initial scan of requirements ahead of any workshop with decision-makers was essential. This should be conducted using a structure that links decision-making needs to formal policy drivers to provide the evidence base. The method adopted in this study was relatively straightforward and also provided an opportunity to merge common stakeholder requirements prior to the workshop to 'clean' the data used on the day. The initial scan also provided policymakers with an opportunity to consult internally around needs prior to any workshop.

Figure 10. Five key priority elicitation phases. Stages 1–4 were conducted during the workshop. Green markers in Stage 3 represent the specific policy needs mapped against their policy/decision importance and feasibility of delivery with ALA/CSIRO data and methods.



- Although the workshop had ambitious objectives that may not have been fully addressed, the breadth of discussion across policy and feasibility provided attendees with an improved understanding of their respective needs and challenges.
- A structured approach to workshop facilitation was essential. This included incorporating structured sessions around policy priorities and feasibility, and gave all participants an opportunity to contribute to the workshop.

5.4.2 What proved challenging

- Assessing feasibility in real-time using a binary scoring system (Yes/No) was challenging as it was only suitable for providing an initial 'coarse-filter' evaluation of feasibility. It was only useful in excluding requirements that were clearly out of scope.
- The approach required improved approaches for more critically and objectively evaluating feasibility. This would include developing definitions around 'feasibility', for example with regard to whether the capability was currently available or whether it was feasible because it could be developed if sufficient resources were available.

- Assessing feasibility via a binary score was limiting where partial solutions were feasible. The workshop provided insufficient time to explore these partial solutions.
- Improved approaches for assessing feasibility that rely less on subjective assessments during a workshop and more on structured approaches are needed.
- The workshop contained one provider actor whereas in reality there would be multiple providers and future priority setting exercises should attempt to integrate multiple provider views. An additional component here was that although the provider may not have all the capability, jointly-delivered capabilities could provide useful solutions.

5.5 Methods recommendations

- There is a requirement to develop simple frameworks to facilitate ongoing discussions around needs and feasibility. As one example, understanding decision-making needs around biodiversity in the context of surveillance monitoring, government program performance monitoring or research monitoring provides a useful construct for better understanding needs. These frameworks are required for both the policy and provider components and will facilitate a clearer dialogue between both stakeholders as terminology and context is well-defined.
- Providing sufficient lead-time to conduct an initial-scan of requirements with policy makers, aggregating this ahead of any joint meeting and potentially socialising joint perspectives prior to any workshop is recommended.
- A more comprehensive initial evaluation of current national capability should be conducted rather than relying on one provider cohort. This could include an inventory of current operational products and services at a general level described against a simple classification to allow for comparability.

5.6 Thematic focus for biodiversity profiling

Workshop-derived priorities and additional lines of evidence, including, for example, key decision-making needs identified in the *Statement of Australian Government Requirements for Environmental Information* published by the Australian Government Environmental Information Advisory Group in 2012 (<http://www.bom.gov.au/environment/about/advisory-group.shtml>) highlighted the need to gain an improved understanding of historical changes in biodiversity composition at continental scales. A possible solution to this challenge is presented in Section 6 by coupling ALA data, readily available time series satellite image derived vegetation metrics and ecological modelling methods appropriate for ecological observations such as those maintained within the ALA. The case study intentionally leverages a suite of operational products available at national scales to explore practical implications of developing such a capability. The case studies focus on historical change rather than on predictions.

6. Case study – biodiversity profiling

6.1 Introduction

There is no single correct answer to the question ‘why, what and how should biodiversity be monitored in Australia?’. Different policy, assessment and planning processes place, and will continue to place, quite different demands on biodiversity monitoring (Ferrier 2012). It is unrealistic to expect that all of these demands will ever be addressed effectively by any one monitoring approach that focuses at a single spatial scale, measures a single dimension and/or level of biological organisation, and employs a single mode of observation.

The case study presented here is intended to illustrate the potential use of ALA data in monitoring, and therefore focuses on a particular monitoring approach to which these data are well suited. It is envisaged that this approach could form one component of a broader integrated program of biodiversity monitoring that would encompass multiple monitoring approaches tailored to address different policy, assessment and planning needs by working with multiple levels and dimensions of biodiversity, at multiple spatial scales, and employing multiple modes of observation.

Figure 11 and Figure 12 provide a conceptual framework indicating where the particular approach illustrated by this case study fits within the bigger picture of biodiversity monitoring. Figure 11 lays out options, or alternatives, relating to the ‘why?’ and ‘what?’ of monitoring, in terms of purpose of assessment, spatial scale, biodiversity dimension, and biodiversity level. The spatial coverage (spread) of the biological datasets accessible through the ALA, and the spatial precision with which individual records are geo-referenced, means that these datasets are generally best suited to applications at a national, or in some cases regional, scale.

Biological datasets accessible through the ALA are, in general, of more relevance to surveillance monitoring, i.e. tracking changes in the overall

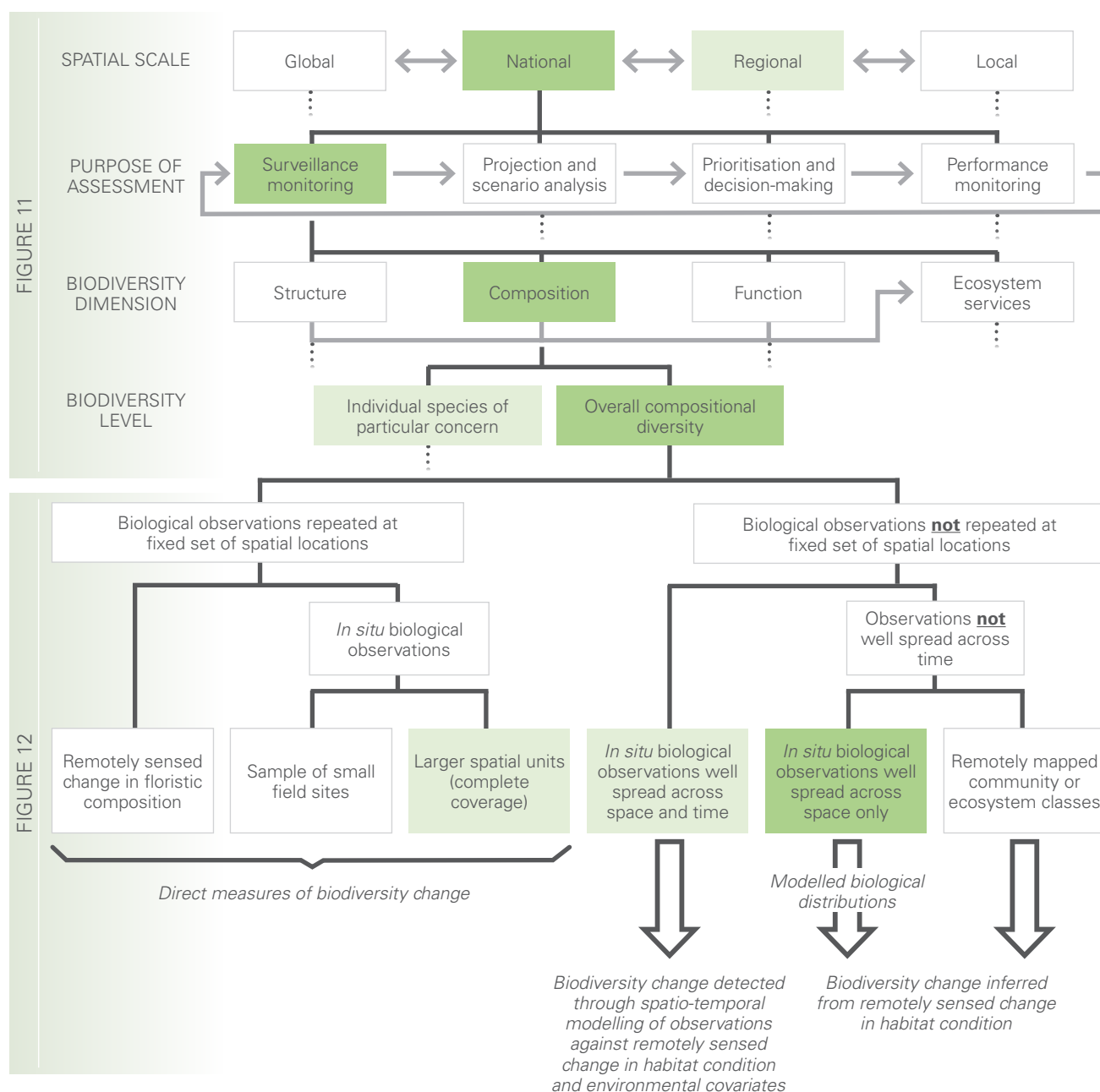
(system-wide) state of biodiversity, than to performance monitoring, i.e. evaluating outcomes for biodiversity resulting from particular policy or management interventions (Figure 11). It should be noted, however, that the particular analytical methodology applied in this case study offers a good potential foundation for fostering strong, and highly dynamic, links between surveillance monitoring and other major biodiversity assessment activities at national and regional scales, including: scenario analysis and projection; prioritisation of, and decision-making around, management actions; and performance monitoring (Williams et al 2010a, b; Ferrier 2012).

ALA data are also, in general, of more relevance to addressing the compositional dimension of biodiversity—i.e. focusing on the identity of, and taxonomic and phylogenetic relationships between, species—than to addressing structural or functional attributes of biodiversity, or whole ecosystems, or to addressing ecosystem services. Within the context of biodiversity monitoring, this compositional perspective can lead either to a focus on assessing change in the state of species of particular conservation concern (e.g. EPBC listed species), or to a broader consideration of change in overall compositional diversity across whole biological groups. While there is considerable potential to make more effective use of ALA data to address the former challenge (focusing on individual species), the case study presented here purposely opts for broader consideration of overall compositional diversity, thereby taking maximum advantage of the remarkably rich taxonomic coverage provided by the ALA datasets.

Figure 12 lays out options relating to the ‘how?’ of biodiversity monitoring, focusing specifically on monitoring approaches of relevance to the particular combination of ‘why?’ and “what?” options highlighted in Figure 11—i.e. national-scale surveillance monitoring of change in overall compositional diversity. A primary division between

Figure 11. A hierarchy of options and alternatives relating to the ‘why?’ and ‘what?’ of biodiversity monitoring. Options highlighted in dark green are those addressed by the case study presented in this section, while options highlighted in light green have good potential to be addressed in the future using ALA data.

Figure 12. Options relating to the ‘how?’ of biodiversity monitoring, focusing specifically on approaches of relevance to national-scale surveillance monitoring of change in overall compositional diversity (from Figure 11). The option highlighted in dark green is the one illustrated by the case study presented in this section, while options highlighted in light green have good potential to be applied in the future using ALA data.



approaches is made here on the basis of whether or not observations of biological composition are repeated (over time) at a fixed set of spatial locations. Of the approaches that do involve repeated biological observation at fixed locations, perhaps the best known are those relying on in situ (field based) observation—i.e. either repeated sampling of relatively small field plots (or transects), such as the networks of ecological monitoring plots being established by TERN (<http://www.tern.org.au/>) or repeated observations aggregated within much larger spatial units, such as the Atlas of Australian Birds (http://www.birddata.com.au/about_atlas.vm). The emergence of new hyperspectral sensors is now also opening up potential to directly monitor change in the floristic composition of canopy vegetation through remote sensing (Wang et al. 2010).

While repeated biological observation at a fixed set of spatial locations is often, quite understandably, viewed as the ‘gold standard’ of biodiversity monitoring, this strategy, if applied on its own will struggle to provide sufficient geographical and taxonomic coverage for some purposes—including the focus here on national-scale surveillance monitoring of change in overall compositional diversity. In such instances it may make good sense to complement, and integrate, this approach with an alternative strategy that combines remotely-sensed change in habitat condition with best-available information on the spatial distribution of biodiversity to infer changes in the retention of compositional diversity (Figure 12) (Ferrier 2011). At the most basic level this second strategy can employ mapped community or ecosystem classes (e.g. NVIS vegetation types <http://www.environment.gov.au/erin/nvis/>) as broad surrogates for biodiversity composition. Major recent advances in the availability of primary biological observations (species occurrence records) and fine-scaled environmental variables for the Australian continent, has now opened up unprecedented potential to model spatial patterns in the distribution of a wide range of biological groups, thereby providing a finer-

scaled, and more explicit, foundation for inferring biodiversity change from remotely-sensed change in habitat condition.

The case study presented here illustrates this latter approach using biological and environmental data accessible through the ALA to model spatial patterns in the distribution of four selected biological groups. This is achieved using generalised dissimilarity modelling (GDM; Ferrier et al. 2007), a technique also now available through the ALA. These modelled patterns are then used as a filter, or lens, through which to infer the consequences, for retention of compositional diversity, of temporal changes in habitat condition nationally, estimated from best-available remote sensing products.

It should be noted that the approach illustrated here uses data only on the distribution of ALA species records across space, not across time. In other words, these records are being used to estimate patterns of distribution in geographic space alone. The subsequent inference of biodiversity change over time in this approach relies completely on the intersection of these spatial patterns with time series data on habitat condition derived from remote sensing. While beyond the scope of the current project, there is considerable future potential to make more effective use of ALA data for biological groups with observations well spread across time, as well as space (but where these observations are largely opportunistic in nature, rather than being repeated observations made at a fixed set of locations) (Figure 12). Specifically there is good potential to extend the GDM-based approach employed below to enable simultaneous modelling of compositional turnover in both space and time, using change in habitat condition (and possibly also change in climate) as covariates, thereby providing a means of extracting signals of biodiversity change from noise due to changing sampling bias.

6.2 Modelling spatial pattern in biodiversity composition

This section outlines the methods used to model the spatial pattern of biodiversity composition using Generalised Dissimilarity Modelling (Ferrier et al. 2007). GDM is a statistical technique for modelling the compositional dissimilarity between pairs of geographical locations for a given biological group as a function of environmental differences between these locations. The 'compositional dissimilarity' between a given pair of locations is most easily thought of as the proportion of species occurring at one location that do not occur at the other location (averaged across the two locations)—ranging from 0 if the two locations have exactly the same species through to 1 if they have no species in common.

GDM uses data on the presence and absence of all species in a group of interest at a sample of locations across the region of interest to fit a model predicting the compositional dissimilarity between pairs of locations as a non-linear multivariate function of the environmental attributes of these locations. Another way of viewing this is that GDM effectively weights and transforms the environmental variables of interest such that distances between locations in this transformed multidimensional environmental space now correlate, as closely as possible, with observed compositional dissimilarities between these same locations (see Ferrier et al. 2007 for full explanation).

Once a GDM model has been fitted using biological data from a sample of locations, this model can then be used to predict the level of compositional dissimilarity expected between locations lacking biological data, based purely on the mapped environmental attributes of these locations. This predictive capacity provides, in turn, a foundation for performing a wide variety of subsequent spatial analyses—e.g. conservation assessments, survey gap analyses, and space for time substitution such as climate change projection. While developed

originally in Australia this general approach is now being explored increasingly in other parts of the world—e.g. in Panama (Faith and Ferrier 2002), Madagascar (Allnutt et al. 2008), New Zealand (Overton et al. 2009) Ecuador (Thomassen et al. 2010), North America (Fitzpatrick et al. 2011) and globally (Ferrier et al. 2004). This community-based method has been applied in marine (Leaper et al. 2011) and freshwater (Leathwick et al. 2011) ecological systems and to a wide range of terrestrial biological groups (e.g., Ashcroft et al. 2010; Burley et al. 2012; Marsh et al. 2010; Rosauer et al. in press; Williams et al. 2012; Williams et al. 2010a).

This process requires the compilation and vetting of biological data, acquisition and development of suitable spatial environmental data, and a systematic process of model fitting. In addition to the prediction of compositional dissimilarity, the outputs of a GDM model are the input environmental layers transformed to best-align with the spatial distribution of records from the assemblages of species. These outputs then form inputs into subsequent analyses, such as those listed above, and for conservation assessment purposes, they may be combined with time series habitat condition layers to evaluate loss or gain in compositional diversity, such as demonstrated through this case study.

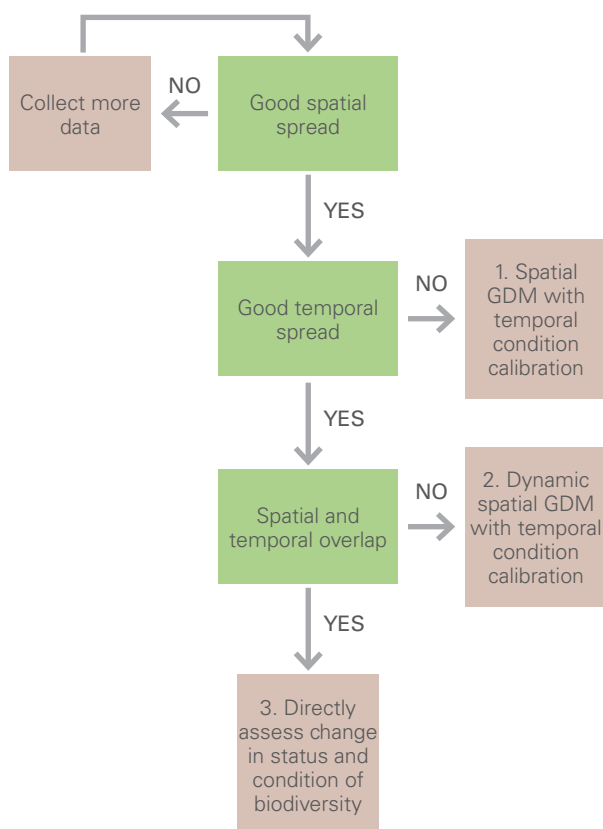
6.2.1 Biological data vetting

The Atlas of Living Australia aggregates biological occurrence records from multiple sources and conducts data quality checks on taxonomic, spatial and temporal information to improve fitness-for-use (<http://code.google.com/p/ala-dataquality/>; Miles 2011). To ensure the suitability of a particular biological group for use with Generalised Dissimilarity Modelling (GDM), additional vetting procedures are needed. For the GDM model to perform well, the biological group should contain a relatively large number of occurrence records (for example, ~> 300 000 for continental Australia) that represent most of the biogeographic regions in

which species from that group are found. A biological group with a relatively large number of species (for example, $\sim > 1000$) is also likely to perform better in GDM than a group with relatively few species. Ideally, the aggregate of records within a grid cell should approximate presence-absence sampling, although this is rarely the case for data sourced from Natural History collections. Large aggregations of data from multiple sources can begin to approach a presence-absence sample at coarse resolutions in some parts of the landscape, but require an assessment of sampling adequacy. Further, vetting procedures were therefore implemented to assess whole assemblage fitness-for-use in GDM at the grid resolution of interest, 0.01 degrees (~ 1 km).

To support more advanced assessment of data fit-for-purpose, an application has been developed for advanced users to permit profiling of biological data, according to specified taxonomic, spatial and temporal parameters (see <http://spatial.ala.org.au/alaspatial/sxs>). The spatial and temporal spread of biological data across the Australian continent, for example, will determine the suitability of three example modelling frameworks (Figure 13). The objective of this advanced profiling tool is to identify the spatial and temporal spread of biological records for different taxonomic groups and capacity to improve these dimensions by grouping records at different spatial and temporal resolutions. The advanced user can investigate the number of species of a particular taxonomic group within unique locations for a specified resolution (e.g. 0.01 degrees) and spatial uncertainty threshold (e.g. $< +/ - 2000$ m) by different spatial (e.g. bioregions) and temporal groupings (e.g. decadal collecting windows). Where we lack adequate temporal biological data but have a good spatial spread, the modelling results can be combined with temporal remote sensing of land cover or land use mapping calibrated with inferred estimates of biological condition. Where we have observations over time in different places, other covariates with a temporal dimension, such as climate variability are needed

Figure 13. A decision framework for evaluating biological records to determine which modelling framework is most applicable.



to make sense of the data, and the results can be combined with temporal remote sensing of land cover or land use mapping calibrated with inferred estimates of biological condition. Where there is good coincidence between the spatial and temporal dimensions of the biological data, then direct modelling of biodiversity status and condition is possible, correlated with remote sensing data and climate variability.

Subsets of biological data can also be generated using a large number of facet variables that are relevant to vetting decisions (see the ALA's Spatial Portal 'Add to map | Facet' tool). Once an assemblage has been defined, it can be viewed using the 'assemblage' option in the 'Add to map | Species'

tool. These tools facilitate use of GDM in the Spatial Portal ('Tools | GDM'). These applications enable the advanced user to further profile the biological data of choice in relation to different facet criteria such as environmental and land use classes or attributes associated with the biological record itself.

To complement the advanced user tools available via the ALA, options for downloading and vetting large volumes of biological data using the R program (<http://cran.r-project.org/>) were explored. Data were extracted directly from the ALA biocache (biocache.ala.org.au) and put through a series of queries to assess fitness-for-use in GDM. The process of data extraction, analysis and mapping was implemented in R scripts. Because data in the ALA are continually updated, the R scripts enabled the vetting process to be repeated at intervals, or when an extraction is to be used in a GDM. This interface for the advanced user is limited to the format of data presented in the biocache. Feedback to the ALA throughout this process resulted in additional flexibility being made available.

Vetting tools were used to evaluate the fitness-for-use of available data for 19 invertebrate groups and for fungi. The study purposely focused on these taxa because they represent highly-diverse components of biodiversity that have been largely neglected by past efforts to assess biodiversity change at a National level. Once the data for each of these groups were downloaded, an initial summary

was run to characterise data quality, as outlined in Table 5. These summary measures compared the proportion of records suitable for analysis relative to all available records. The vetting process considered the taxonomic identification level, the typical habitat as terrestrial, marine or aquatic (freshwater), spatial accuracy and date of occurrence (Table 6). Only data vetted at the species level was used because information at the subspecies level is inconsistently reported across multiple sources. The approach to vetting is summarised in four tables that build on each stage of the process (Table 7). Records for each biological group are initially vetted to the species level, and then analysed for spatial accuracy, and date accuracy. Finally, a synthesis table compares the results across biological groups using ranks (for example, 1–19 invertebrate groups evaluated). At each stage of vetting, each group was ranked based on the results in each data quality category. Groups that scored high ranks across a number of categories were indicative of better prospects for GDM modelling. As spatial accuracy is only sparsely annotated, biological records that lack spatial accuracy were included in the data compiled for GDM analysis. The majority of groups had some date associated with the records. This limited process made good use of preliminary vetting by the ALA.

Based on this vetting, four groups were selected for further consideration in developing fitted GDM models for the case study: Suborder Apocrita (bees, wasps and ants), Suborder Araneae (spiders),

Table 5. Basic data profile attributes.

Data profile description	Details
Total number of unique records (n)	Measured by latitude and longitude rounded to the analysis resolution (in our case, 0.01 decimal degrees).
Total number of unique species (n)	
Geographic spread of the data	Measured by number of IBRA 6.1 bioregions (of 85 continent-wide) in which occurrence points were located.
Density of unique locations = # unique (Lat/long/species) / # records * 100	Measured using the number of unique latitude, longitude, and species combinations compared to the total number of records to determine the proportion of records which were unique.

Table 6. Key components of the vetting process.

Vetting description	Details
Identification level – matched species	Occurrence records matched to species level. This includes subspecies identifications which are assigned to species level and manuscript names, but excludes indeterminate species (i.e., sp., ?, cf, aff, hybrids denoted by 'x').
Life history habitat	Marine and fully aquatic species were excluded. Species with a land phase in their lifecycle were included, noted by the interim register of marine and non-marine genera http://www.cmar.csiro.au/datacentre/irmng/ , an authoritative source used by the ALA.
Life history habitat	Each occurrence record was assigned to one of 6 categories: no spatial accuracy defined and spatial accuracy in 5 classes: ≤ 250 m, ≤ 1000 , ≤ 5000 , ≤ 10000 , $>10\ 000$ m.
Date of observation	Records were assigned to the following date ranges: pre 1950 (< 1950), between 1950 and 1975 (≥ 1950 and < 1975), post 1975 (≥ 1975).

Table 7. Tabulation of results (see authors for details).

Table description	Details
Records vetted to species level	Table 1: Summary of all records, records vetted to species level and proportion of records vetted to species level. The number of records, number of species, number of IBRA and number of unique latitude/longitude (site).
Records vetted to species level and spatial accuracy	Table 2: Summary of the per cent of records by given spatial resolution for each taxonomic group of interest.
Vetting by date accuracy	Table 3: Summary of taxon groups (vetted to species level) with number of records per date category (NA date, Pre 1950, 1950–1975, Post 1975) and proportion of total records (%) with No date or Post 1950.
Comparison across groups	Table 4: Overall group ranking based on vetting. Groups range from 1–19 with 1 being the lowest and 19 being highest ranked within each category. Ranking were assigned for: • Total number of records (identified to species level) • Total number of species (identified to species level) • Total number of unique locations (latitudes and longitudes) • Total number of IBRA regions (to determine the spread across the country) • Density = No of unique Lat/long/species / No of records * 100 (%). This provides a measure of the proportion of records which are unique lat/long/ species. • Number of records with spatial resolution (i.e. NB this can be up to 10 000 m) • Number of records with dates • Number of records with dates (post 1950).
Sampling adequacy	A series of maps showing the geographic spread of all vetted records by the number of species within a grid cell to determine whether filtering on richness as a proxy for sampling adequacy is needed, e.g. singletons (1 species), ≥ 1 , ≥ 2 , ≥ 3 , ≥ 5 , ≥ 10 species (see example Figure 14).

Figure 14. Example maps for Kingdom Fungi showing the geographic spread of vetted records satisfying different thresholds for the number of species in a grid cell.

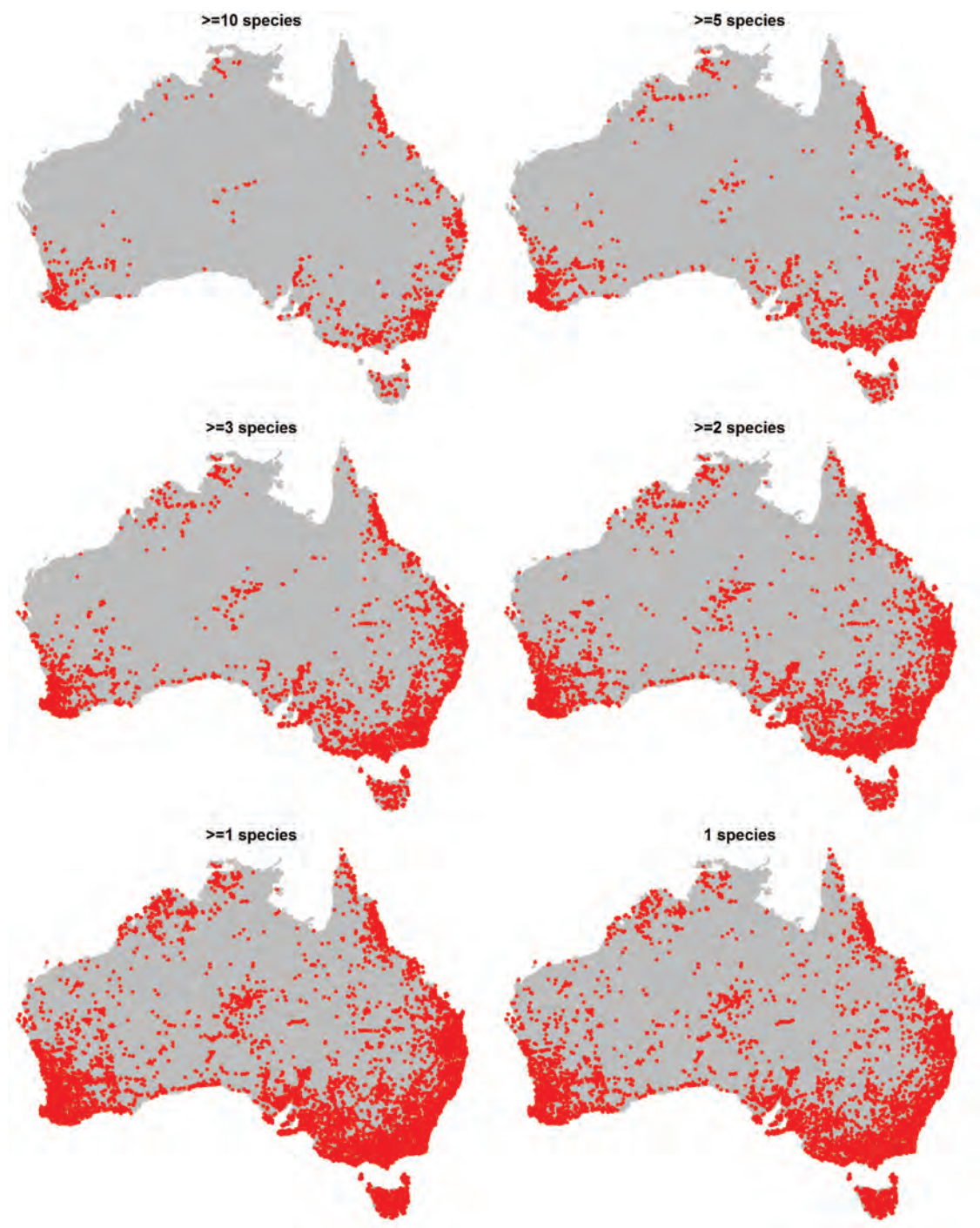


Table 8. Fitted GDM models.

Biological group	# 0.01 grids	# species	# site-pairs fitted	Fitted variables (see http://spatial.ala.org.au/layers for abbreviated layer names)
Fungi	16 976	3630	900 000	CLIMATE (13 variables): rainx, raini, radni, microgi, mesogi, megagi, maxti, evapi, adefx, srain2, srain1, rtxmin, rprecmin SUBSTRATE (8 variables): slope, roughness, magnetics, nutrients, pc3_80, pc3_20, clay30, bd30
Apocrita	9867	3819	900 000	CLIMATE (11 variables): raini, radnx, minti, microgi, mesogi, evapi, arid_min, adefi, srain1, rprecmin, rpremax SUBSTRATE (9 variables): twi, roughness, magnetics, gravity, nutrients, pc3_20, kao80, pawc1m, bd30
Aranae	9734	2192	900 000	CLIMATE (9 variables): radni, mesogi, megagi, c4gi, arid_min, arid_max, adefi, rtimin, rpremax SUBSTRATE (8 variables): slope, magnetics, geollrngeage, wiioz2, nutrients, pc1_20, kao20, clay30 Geographic distance predictor included
Coleoptera	13 350	9248	850 000	CLIMATE (11 variables): raini, radni, microgi, mesogi, arid_min, adefi, trngi, srain2, rtxmin, timin, rprecmin SUBSTRATE (7 variables): slope, ridgetopflat, magnetics, nutrients, kao20, ill80, bd30 Geographic distance predictor included

Order Coleoptera (beetles) and the Kingdom Fungi (mushrooms, moulds and yeasts). Data for these groups were further mapped to assess their geographic spread or spatial bias. To test for issues of sampling adequacy, that may depress the fit of the GDM model, separate maps are produced for different numbers of species aggregated to a grid cell. Figure 14, for example, shows the tradeoff between the number of species recorded at a site and the geographic spread of the data for the Kingdom Fungi. As the threshold number of species at a site to include in the GDM analysis is increased, as a surrogate for improved within-site sampling adequacy, the spatial adequacy of sampling between sites declines. This analysis helps in making a decision about the threshold of species richness to use (e.g. > 2 species) for a given spatial resolution

(e.g. 0.01 grid) as filter applied to the biological data used in GDM. Alternatively, or in addition, a weight variable derived from the sum of the number of species at two sites in a pair used to calculate dissimilarity can be applied with the response when statistically fitting the GDM model to the data (see Ferrier et al. 2007 for details).

6.2.2 Environmental layers

The environmental layers (variables) used to model spatial pattern in biodiversity composition in this case study, were drawn from a consistent set of 0.01 degree (~1 km) gridded variables developed by CSIRO for GDM modeling (Williams et al. 2010a, b; Williams, Belbin et al. 2012), and provided to the ALA for demonstration purposes. These layers can

be distinguished from other ALA environmental layers by the key word 'GDM' when viewed at <http://spatial.ala.org.au/layers>. To enable high resolution modelling in local regions, a set of 250 m gridded environmental variables have also been developed by CSIRO and provided to the ALA.

6.2.3 Model fitting

A demonstration version of the GDM software is available on the ALA's Spatial Portal. Help for the application is available at <http://www.ala.org.au/spatial-portal-help/gdm/>. Models were fitted using the stand-alone .NET version of this software written by Glenn Manion, GD Modeller version 2.02, applied to data downloaded from the ALA biocache, as described above, and 1 km gridded environmental variables (detailed in Williams et al. 2010a, b). To select the most important variables for inclusion in each model, we followed the model fitting approach outlined by Williams, Belbin et al. (2012). Correlated sets of environmental variables (31 climate, 19 soil, six geoscience, eight terrain) were initially fitted and then those variables that contributed to the fit were combined in a single model. A backward elimination procedure removed correlated variables that contributed little to the model. A stopping point was reached when the reduction in percent deviance explained exceeded 0.03, typically resulting in retention of around 20 variables. This exploratory model fitting used a sample of 300 000 site pairs. The final model was fitted used 900 000 site pairs. Site pairs were randomly selected from the very large number of possible pairs (2^n). A test for residual variation explained by geographic distance was conducted on the final model. Geographic distance was included as a predictor if it significantly (percent deviance explained > 0.03) and independently (without dropping other variables) improved the model.

As mentioned above, the primary outputs from GDM are the input environmental layers transformed to best-align with the spatial distribution of records

Figure 15. Example classification derived from GDM modelling of spatial pattern in biodiversity composition for Kingdom Fungi. Areas with similar colours have similar groups of fungi. The GDM analysis defines and scales the environmental factors that distinguish these areas.



from the assemblages of species. These layers can be used in a wide range of post-processing analyses. One of these analyses, an environmental domain classification, is available on the ALA's Spatial Portal. Example classifications using an alternative hierarchical clustering method (UPGMA) with 400 classes, is shown in Figure 15. The next section demonstrates how these outputs from GDM can be combined with time series vegetation condition layers to evaluate loss or gain in compositional diversity.

6.3 Continental-scale measures of habitat condition

Information about change in condition of native vegetation or the quality of habitat to support biodiversity is fundamental to effective management of land for conservation values. Habitat condition can be measured at the site level and more broadly using remote sensing detection, calibrated by field observations and relevant data sources (Briggs and Freudenberger 2006). However, reliable measurement or modelling of habitat condition

dynamics beyond simple measures of vegetation cover and extent has eluded mainstream reporting (for examples, see Zerger et al. 2006, 2009; Drielsma and Ferrier 2006; Cook et al. 2010).

In the absence of comprehensive, integrated approaches to vegetation condition assessment, dynamic land use and land cover mapping can provide an interim measure of change in habitat condition. A simple state-transition framework for classifying vegetation condition, for example, was developed by Thackway and Lesslie (2006) and applied nationally by integrating a wide range of land cover and land use datasets to provide an indicative condition baseline (Thackway and Lesslie 2008). A simple intuitive scaling of biodiversity response to disturbance from land use intensity ranks was developed by Williams, Reeson et al. (2012) and benchmarked against the framework of Thackway and Lesslie (2006) and the state-transition modelling of Drielsma and Ferrier (2006). Inferred vegetation condition was used to assess effective habitat areas. As knowledge accrues, such intuitive scaling can be replaced by calibration models, and implemented dynamically where time series data are available.

The purpose of this section is to outline the range of time series continental-scale land use and land cover mapping and remote sensing data that are currently available and have potential application to dynamic habitat condition mapping. Other datasets such as 3D photography, hyperspectral imaging and Lidar are not discussed here as they are presently not consistently available across the entire Australian continent for public use. As outlined in the preceding sections, time series condition data can be combined with biological data to model temporal change in biodiversity composition. Key datasets, their characteristics and limitations are summarised in Table 9, at the end of this section.

6.3.1 Remotely-sensed vegetation indices

Spatial information about the green foliage cover of woody evergreen vegetation can be obtained through the analysis of time series of satellite imagery (Berry and Roderick 2002a, b; Roderick et al. 1999). Berry and Roderick (2002) demonstrated a method for decomposing the total amount of photosynthetically active radiation absorbed by the vegetation (fPAR) into a mixture of three basic functional leaf types—‘turgor’ (T), ‘mesic’ (M), and ‘sclerophyll’ (S)—by modelling also the relationship with solar radiation and rainfall. In Australia, where most trees and shrubs are evergreen, the T component is comprised almost entirely of ephemeral herbaceous species. This ternary ‘TMS’ framework has been used to map changes in Australian vegetation in response to land use and climate (Donohue et al. 2009; Berry and Roderick 2002, 2006). With reference sites to calibrate potential habitat quality and relative naturalness, it has potential application for vegetation condition monitoring. The approach can be applied to remotely derived photosynthetically active radiation from greenness or vegetation indices (for example, MODIS and AVHRR).

6.3.2 National land use

National land use is created using a modelled approach to couple agricultural census commodity statistics to satellite imagery such as Advanced Very High Resolution Radiometer (AVHRR) and other sources of land use information. The mapping procedure matches the growth characteristics of different crops and pastures over a one-year period to a Normalised Difference Vegetation Index (NDVI) time series (Bureau of Rural Sciences 2004). Each national dataset comprises a set of probability surfaces for agricultural land uses and a summary categorical land-use map for both agricultural and non agricultural land use. Each national land-use map represents a land-use snapshot at a particular point in time and should be used nominally at 1:2 500 000.

The spatial distribution of land use is modelled and therefore pixel by pixel comparisons should not be made between two temporally different national land-use maps to infer that real land use changes have affected those specific pixels (Australian Bureau of Agricultural and Resource Economics and Sciences 2011). Changes in land use can be identified at a statistical local area level or greater. Each national-scale land-use map is a snapshot of land use at a particular time—the maps do not show cyclical changes in land use, such as crop rotations.

The resolution of AVHRR satellite data is 1 km², which is too coarse to map land use covering small areas, especially where more than one land use occurs within a 1 km × 1 km pixel grid. This can occur where different commodity groups are close in space such as strip cropping and small-scale planting. Attribute accuracy is generally dependant on how distinct the commodity appears in the satellite image (Bureau of Rural Sciences 2010). Some agricultural land uses and crop types are impossible to distinguish with satellite data alone.

6.3.3 Catchment-scale land use

Catchment-scale land-use data are generated by combining the cadastre, public land databases satellite data, such as Landsat or SPOT and land cover and land use information collected in the field (Australian Bureau of Agricultural and Resource Economics and Sciences 2011; Bureau of Rural Sciences 2011). The scale for catchment land use ranges from 1:25 000 for more intensive uses such as peri-urban areas through to 1:100 000 for broad acre cropping regions and 1:250 000 for semi-air and pastoral areas. Catchment-scale land-use data are created by State and Territory agencies in collaboration with ABARES. ABARES collate catchment-scale data provided by State and Territory agencies to produce an Australia-wide catchment-scale land-use map. Updates are sporadic, and as a result, catchment-scale land-use maps for Australia are comprised of data from a variety of different

time periods. The data inputs range from 1997 to the present. For this reason catchment-scale land-use data are not considered a viable option for national habitat condition assessment and biodiversity modelling.

6.3.4 Dynamic Land Cover (v1)

The National Dynamic Land Cover dataset is generated by Geoscience Australia and the Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES), using Moderate Resolution Imaging Spectroradiometer (MODIS) enhanced vegetation index data from 2000 to 2008, consisting of 186 layers of 16-day enhanced vegetation index composites. Every pixel, which is 250 m² is represented by a time series made up of 186 observations (Lymburner et al. 2011). The time series was analysed and described in terms of 12 time-related coefficients that represent the vegetation phenological character and capture seasonal characteristics of the vegetation dynamics. A clustering approach was applied to the 12 coefficients to define homogenous classes that had similar greenness dynamics over time. Regions sharing similar time series coefficients were labelled using information derived from Catchment Scale Land Use (ABARES, 2011) and Native Vegetation Information System datasets (Department of the Environment and Water Resources 2007). Land cover clusters were then labelled (Mutendeuzi and Stafford-Bell 2011). More frequent updates for Dynamic Land Cover are in the development stage and were unavailable for use in this project. Updates are expected to be available for beta testing in the near future (A. MacIntyre pers. comm. 2012).

6.3.5 Forest extent

The National Carbon Accounting System (NCAS) accounts for carbon emissions from land-based activities to meet national and international reporting requirements. The Department of Climate Change and Energy Efficiency used Landsat satellite

Table 9. Comparison of data to infer habitat change or vegetation condition at continental scales.

	National Land Use	NCAS Woody /Non-Woody	Dynamic Land Cover
Summary	Datasets are a time series of national-scale land-use maps. Each dataset comprises a set of probability surfaces for non-agricultural land uses and a summary categorical land use map for both agricultural and non-agricultural land uses.	Landsat satellite imagery was used to discriminate between forest and non-forest cover. Forest is defined as vegetation with a minimum 20% canopy cover, potential reaching 2 m high.	The first complete land cover classification for continental Australia. The dynamic land cover map comprises six hierarchies and 34 classes. The dynamic land cover database is a time series database based on an analysis of 16 day Enhanced Vegetation Index composites for the period 2000–08.
Update frequency	5 yrs – 1992–93, 1993–94, 1996–97, 1998–99, 2000–01, 2001–02 and 2005–06	Epochs are: 1972, 1977, 1980, 1985, 1988, 1989, 1991, 1992, 1995, 1998, 2000, 2002, 2004, 2005, 2006, 2007, 2008, 2009, 2010 & 2011	Potentially annual
Classification	ALUM 6 primary classes 31 secondary classes	2 classes (forest and non-forest)	ISO land cover standard (19144–2) 34 classes
Resolution	1 km	25 m	250 m
Satellite	AVHRR	Landsat	Modis – method is sensor independent
Extent	Australia	Australia	Australia
Attribute	Probability surfaces for agricultural land use and categorical land use map for agricultural and non agricultural land uses.	Forest and non-forest.	Land cover information based on temporal behaviour of every 250 m pixel from 2000–08
Limitations	Not suitable for pixel by pixel comparisons. Comparisons suitable at a statistical local area level or greater. Land uses less than 1 km not identified Attribute accuracy dependant on how distinct a commodity appears in the satellite image Some commodities are impossible to distinguish using remote sensing.	Limited classification restricted to forest and non-forest	Annual updates are in development. They are not currently available.

Source: Adapted from Mutendeudzi and Stafford-Bell 2011.

imagery to discriminate between forest and non-forest cover. Forest is defined as vegetation with a minimum 20 per cent canopy cover, potentially reaching two metres high and a minimum area of 0.2 hectares (DCCEE 2112). Forest cover and change maps (deforestation and forest conversion and regrowth) are produced at a 25 m pixel resolution for the following epochs: 1972, 1977, 1980, 1985, 1988, 1989, 1991, 1992, 1995, 1998, 2000, 2002, 2004, 2005, 2006, 2007, 2008, 2009, 2010 and 2011 (DCCEE 2112).

6.4 Estimating change in vegetation condition through remote sensing

The development of a series of reliable spatial layers showing past and present change in vegetation (or ecosystem) condition is an ongoing problem in Australia and beyond (see Section 6.3). While some time series condition layers were developed as part of the Caring for Our Country project (Williams et al. 2010a, b), these were based on the national-level land use data (Bureau of Rural Sciences 2006, 2009) which have some known temporal inconsistencies (see Table 9). An approach was investigated for the case study presented here to infer condition based on the National Carbon Accounting System (NCAS) Forest Extent dataset (DCCEE 2012). This dataset is useful for detecting change only for woody vegetation types (see discussion of limitations in Section 6.3). In this dataset, woody vegetation is defined as vegetation with a minimum 20 per cent crown cover (about 12 per cent foliage projective cover), potentially reaching two metres high and a minimum area of 0.2 hectares (Furby 2002). However, since the time period covered by this dataset starts 30 years ago, and is expected to be updated at regular intervals, these data potentially provide a good source of information for estimating change in biodiversity associated with woody vegetation.

6.4.1 Input datasets

Four datasets were used to derive the time series of woody-vegetation condition employed in this case study.

- The National Carbon Accounting System (NCAS) Forest Extent dataset version 8 (DCCEE, 2012). This nine second (250 m) resolution raster dataset holds remotely sensed forest cover data for a large proportion of the continent, excluding non-forested areas, for 1972, 1977, 1980, 1985, 1988, 1989, 1991, 1992, 1995, 1998, 2000, 2002, 2004, 2006, 2008. The data values are the area in hectares of forest cover detected at a 25-metre resolution from Landsat MSS, TM and ETM+ imagery.
- The estimated original (pre-European) distribution of Major Vegetation Groups from the National Vegetation Information System version 3.1 (DEWR 2007) at a 0.01 degree (1 km) resolution from which an ideal Forest Cover per nine second cell was determined (i.e. the cover expected for this cell if it were in pristine condition).
- The CAPAD_08 (NRS 2010) National Reserve dataset was used to identify areas in good (baseline) condition for each Major Vegetation Group.
- The Australian Plantations 2006 dataset (BRS, 2006) was used to mask out plantation forests from the analysis.

6.4.2 Analytical method

The NCAS data for 1972, 1977, and 1980 were averaged to provide a baseline forest cover state for each nine-second grid cell. Since the pre-2000 data did not include a Landsat image in western Queensland, this area was removed from all post-2000 data for consistency. The Australia's Plantations dataset was used to exclude commercial plantations from the analysis. The CAPAD_08 dataset was rasterised to nine seconds, adjacent reserved cells were combined, and then the boundary eroded by one cell to define a core reserve area without

edge cells. It was assumed that cells falling into this core reserve area are of good condition. For all cells within the core reserve area, the mean baseline forest cover in each selected NVIS major vegetation group (MVG) was calculated. It was then assumed that cells with a forest cover equal to, or greater than, this mean are in pristine condition (1), and cover below the mean has linearly declining condition between the mean and no forest cover, such that no forest cover has a condition score of 0. Two criteria were used to assess the appropriateness of the approach for each NVIS MVG. Firstly, the degree of coverage of the MVG by non-zero NCAS data was assessed, both overall and within the core reserves used for calibration. Where the proportion of non-zero cells was greater than 60 per cent within the core reserves, and the overall pattern of non-forest cells was consistent with anthropogenic clearing (for example, Eucalypt Woodland) the MVG was included in subsequent steps of the analysis. Secondly, the suitability of forest cover as a metric for the assessment of condition was examined for each MVG. As forest types become more open, the absence of forest cells becomes a neutral or positive indicator. Only those MVGs which had an average forest cover above 50 per cent were included in the subsequent analysis.

The above approach was applied (using bespoke software developed by CSIRO) to all cells in NVIS MVG types with at least some tree cover, generating a 0 to 1 condition surface for NCAS surveyed woody MVG types for each of the study years at 9 second resolution. This was then resampled to 0.01 degree resolution as the mean of all sub cells, using ArcMap 10.

6.4.3 Outputs

The primary output is a 0.01° (~1 km) condition grid for NCAS surveyed woody MVGs for each of four study years (1975, 1985, 1995, 2006). Due to limitations inherent with NCAS data (Table 9), only the following NVIS MVGs were considered

in subsequent stages of the analysis (based on the criteria specified above under Methods): Rainforests and Vine Thickets, Eucalypt Tall Open Forests, Eucalypt Open Forests, Eucalypt Low Open Forests, Eucalypt Woodlands, Callitris Forests and Woodlands, Low Closed Forests and Tall Closed Shrublands, Heathlands, Mangroves.

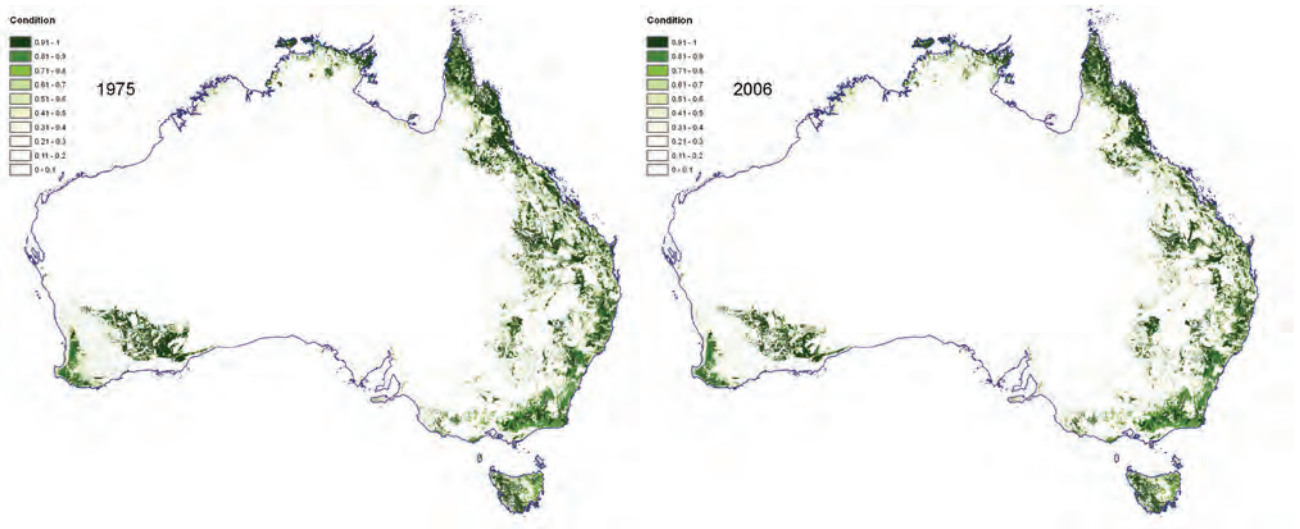
Note that the condition scores are only valid for the forested area within a vegetation type. Due to the nature of the subsequent analysis, the stable condition scores for non-forested grid cells were effectively neutral, so this was an issue for interpretation only. Figure 16 shows the start and end points of the woody-vegetation condition layers time series developed for this case study. Areas without forest cover have a condition score of 0, ensuring that they are neutral in subsequent analysis. These condition grids are then combined with the results of the fitted GDM models to infer biodiversity compositional change attributed to temporal change in forest cover for each biological group, as detailed in section 6.5.

6.5 Inference of biodiversity change

6.5.1 Analytical method

The transformed environmental grids generated by the GDM models fitted to the ALA aggregated data can be used to predict the compositional similarity, s_{ij} , expected between any two 0.01° grid cells (i and j) if those cells were in a natural state. The actual condition c_i of any cell at any time point is estimated in the NCAS condition grids derived above. The proportion (p_i) of species associated with any given cell i (if that cell were in a natural state) that are expected to be retained anywhere within a 300 km radius of this cell, was then calculated by applying a simple species-area relationship to the proportion of similar habitat (to the cell of interest) remaining within that radius (see Ferrier et al. 2004, and Allnutt et al. 2008 for details):

Figure 16. Derived woody-vegetation condition layers for 1975 and 2006 (begin/end points of the time series), with darker greens indicating better condition. While most changes are subtle over this period, a visible reduction in wooded vegetation is evident for central eastern Australia. These condition grids are used with the results of the fitted GDM models to infer biodiversity compositional change (see section 6.5).



$$p_i = \left[\frac{\sum_{j=1}^n s_{ij} c_j}{\sum_{j=1}^n s_{ij}} \right]^{0.25} \quad (1)$$

In this case, in order to limit the edge effects of the forested area modelled in the condition time series, each cell i was compared with all cells j within a 300 km radius.

Two of the four modelled biological groups, Fungi and Coleoptera, were subjected to this final stage of the analysis, assuming constant abiotic environmental conditions (such as climate and soils) but changing vegetation condition, across four points in time: 1975, 1985, 1995 and 2006.

To map change over time, the difference in the proportion of similar habitat (to the cell of interest) remaining at these two time points (t_1 and t_2) was calculated for each cell as:

$$change_i^{t_1-t_2} = \left[\frac{\sum_{j=1}^n s_{ij} c_j^{t_2}}{\sum_{j=1}^n s_{ij}} \right] - \left[\frac{\sum_{j=1}^n s_{ij} c_j^{t_1}}{\sum_{j=1}^n s_{ij}} \right] \quad (2)$$

The methodology described by Ferrier et al. (2004) and Allnutt et al. (2008) also allows estimation of the proportion of species expected to be retained in any defined region of interest, as a weighted average of the p_i values for individual grid cells within that region (with weights reflecting relative distinctiveness in the composition of cells; see these papers for further detail). This technique was applied to five IBRA bioregions (Figure 4 and 5), selected to provide a good spatial spread of results for regions well represented by the forest condition index. While IBRA regions were used here to illustrate this capability, the approach can be used to report change in relation to any other defined area at any scale (for example, a NRM region, a State, or the entire continent).

6.5.2 Outputs

In most parts of the continent, the condition time series showed a reduction in condition over time, and this was reflected in the inferred changes in compositional diversity. As an initial illustration, the outputs of equation 2 are mapped in Figure

Figure 17. Change in retention of compositional diversity within selected woody NVIS MVGs over the past 30 years for beetles (left) and fungi (right). Red-brown colours indicate reduction, yellows indicate no change and greens an improvement. All excluded NVIS categories are shown in grey, and the selected IBRA regions for further analysis highlighted.

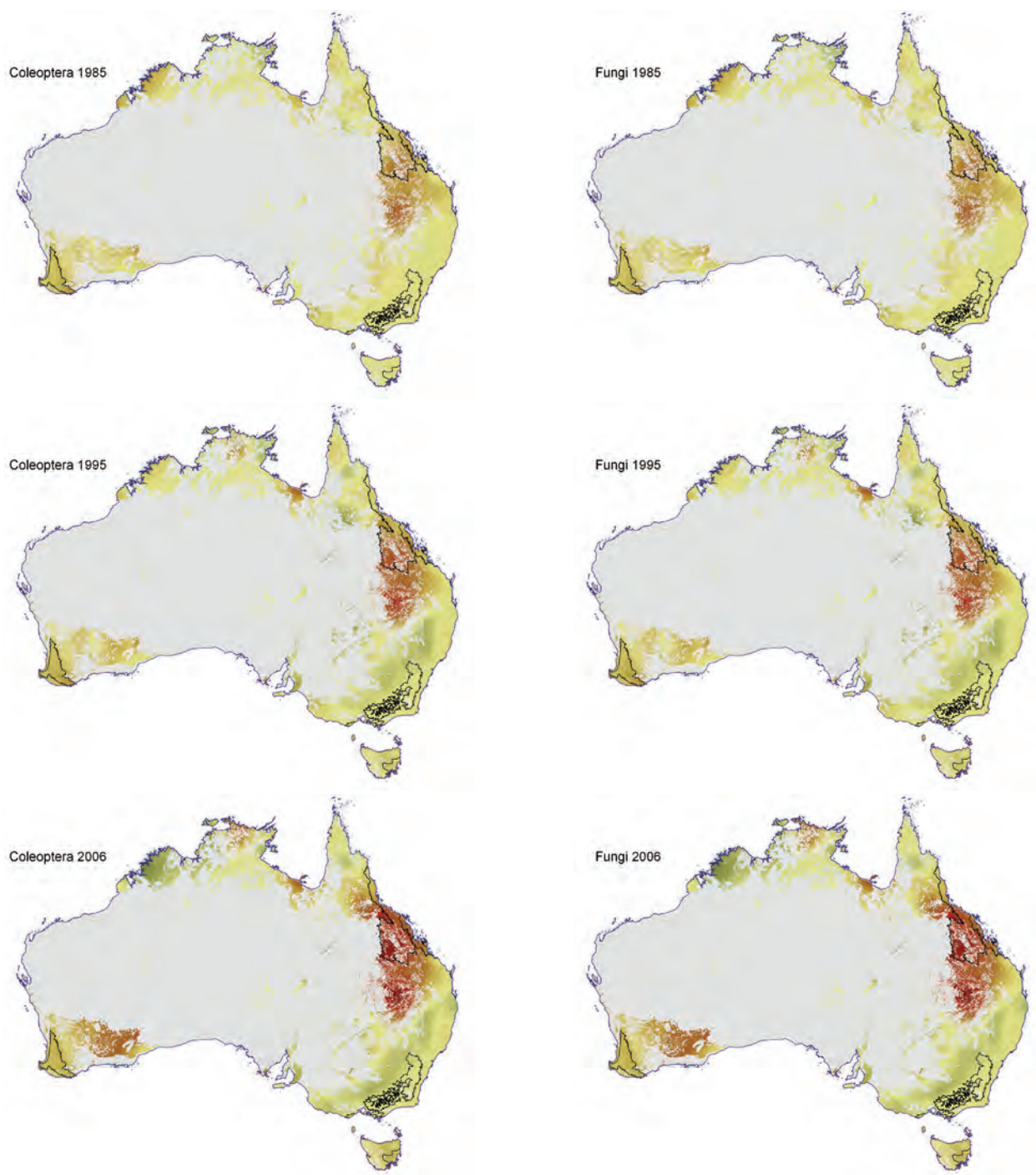
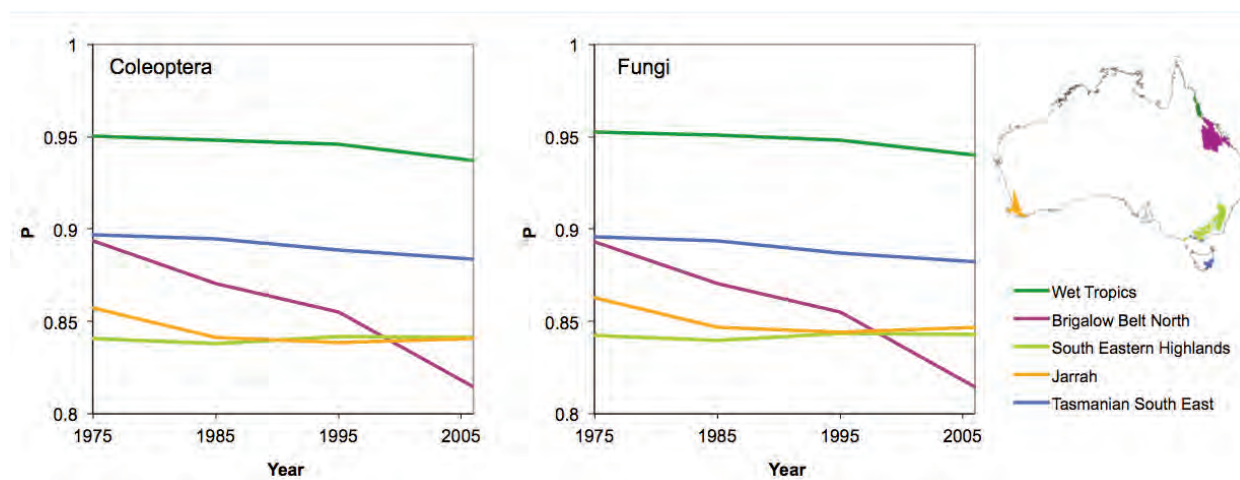


Figure 18. Time series of aggregated (weighted average) p values for selected IBRA bioregions, indicating the proportion of compositional (species) diversity expected to be retained in each of these regions.



17 for both biological groups. The reduction in condition shown in Figure 14 is immediately evident in the brown area to the northeast, as are local improvements in the southeast. However, due to the nature of the condition layers, care must be taken in interpretation. While the Kimberly region shows an improvement over time, this is an improvement in the forested cells only which represents a small proportion of the region.

Changes in the aggregated (weighted average) p values for the five selected bioregions are plotted in Figure 18. They show a pattern consistent with the results mapped in Figure 17. Some differences between groups is evident, but regional differences dominate. Recent clearing in the Brigalow Belt of eastern Australia is reflected in a downward trend, whilst the effects of revegetation efforts in the South Eastern Highlands are apparent.

6.6 Conclusion

The data aggregated by the ALA were readily obtained and filtered to improve data quality as an input, along with spatial environmental data to develop fitted GDM models. The derived forest condition time series has limitations regarding its use because it addresses only woody vegetation condition as defined by the NVIS data (> 20 per cent crown cover, about 12 per cent foliage projective cover). Non-native woody vegetation needs to be better identified and removed, but otherwise the data used resulted in a consistently plausible condition metric. It was a relatively straightforward exercise to derive the biodiversity time series, as a demonstration of the utility of using remotely sensed time series condition data as a basis for continental-scale biodiversity monitoring. Further work is needed to validate the method used to calibrate native vegetation condition and to verify the distribution of native and non-native forest vegetation detected by the NCAS dataset.

7. Conclusion

The preceding sections have explored the potential for leveraging the rich biological data held within the Atlas of Living Australia with modelling approaches that incorporate explanatory spatial data that are typically available at continental scales. The study focussed on data available from operational systems maintained by the Australian government and other key data providers. The use of readily available operational data, including both biological and spatial, was critical in this study to ensure lessons learnt were grounded in this operational context. It is thus important to note that the study should not be seen as a detailed research undertaking but rather an attempt to couple existing data, tested and peer reviewed modelling methods, with spatially explicit explanatory variables to explore options for developing an operational biodiversity modelling capability to inform monitoring.

In addition to the challenges associated with data assimilation, ecological modelling and prediction, the study also explored the role of 'soft' enablers for supporting such an operational capability. Section 3 provided a detailed exploration of these enablers using learnings from the development of the ALA. Although the report focused on the ALA experience, the treatment of enablers provides useful lessons into other domains, for example, soils, earth science, land use, land management and other data that can only be practically acquired by aggregating from many data custodians. This is one of the key contributions of the study. The following provides a summary of conclusions and recommendations derived from the study.

1. Coupling observational data such as the ALA collections with spatially continuous and temporally dynamic explanatory data (such as NCAS satellite imagery already collated for other purposes) provides a sensible strategy for overcoming the relative paucity of temporal biodiversity data in Australia. In other words, establishing strong and ecologically plausible empirical relationships between sparse observational data with those things measured at coarser continental scales is a sensible strategy for understanding change in biodiversity at scales that are useful to support national-scale decision-making.
2. New opportunities for data assimilation are rapidly emerging allowing practitioners to integrate data across domains to obtain new insights into ecosystem change, interactions and potential outlooks. This is increasingly becoming possible at continental scales owing to major capability investments such as the Atlas of Living Australia and the Terrestrial Ecosystems Research Network. Similarly, national environmental data capabilities maintained by Australian government departments such as the Bureau of Meteorology (Climate and Water), Geoscience Australia (Earth Sciences and National Mapping) the Department of Innovation, Industry, Science, Research and Tertiary Education, and the Department of Sustainability, Environment, Water, Population and Communities are examples that provide a rich suite of additional data that can be leveraged to complement any analysis.
 - a. A key challenge in coupling such novel data streams is twofold. The first requires the formation of interdisciplinary teams, for example, ecologists and landscape modellers, to inform the development of suitable analysis methods and provide sensible interpretations of both input and output data.
 - b. The second is in regard to the development of appropriate robust modelling methods and computational resources that can effectively exploit this breadth of spatial and temporal data.
3. The study explored a specific case study based on data held by the Atlas of Living Australia. Opportunities exist for better integrating other observational data such as that also maintained in operational settings by agencies such as SEWPaC and missing State-based data.

4. The establishment of the Atlas of Living Australia has highlighted the importance of establishing a trusted central authority for coordinating the specification and implementation of standards to aggregate and disseminate environmental data from data custodians. Indeed, in some relatively complex domains such as the ecological sciences, a federated model for data acquisition is a sensible strategy. The ALA's success has been partly due to two principles: supporting the sources and custodians of data, and making aggregated data as highly useable as possible through the addition of elegant data discovery, visualisation and analytical tools.
5. The role of standards is critical for supporting both rapid data aggregation and assimilation, but also data dissemination and re-use. The biodiversity profiling case study that was developed in a relatively rapid timeframe was only made possible through significant investment in development and application of standards. This is particularly critical for relatively complex observational data such as ecological data.
6. Aggregating data from multiple providers will need to adopt relatively flexible information architectures including both centralised and distributed models. Owing to the relatively static nature of ecological data and the relatively large number of records required to provide data at continental scales, centralised models are likely to remain the only workable option for analytical infrastructure. The central repository mirrors the data held by custodians, but offer significant efficiencies in regard to data retrieval, interrogation, analyses and publication.
7. Metadata play a critical role in supporting the effective use and re-use of environmental data, particularly in regard to enabling stakeholders to assess data for fitness-for-use. This is particularly true in institutionally dynamic settings where lack of corporate knowledge may hinder effective data use and re-use. Capturing quality metadata at time of data capture cannot be overvalued.
8. Establishing a culture of continuous environmental data availability and free and open access amongst multiple stakeholders is likely to provide new opportunities for supporting more informed decision-making. However, to enable this requires policy and or legislative support, for example through the establishment of open access policies by agencies. Open data principles are currently being championed by the Office of the Australian Information Commissioner and related initiatives such as AusGOAL and these will be key to achieving uptake.
9. Efficient sharing of data is very difficult unless a common licensing framework is adopted for access and re-use of data. It is the ability to combine data from different sets and contributors that really provides value to the users of aggregates.
10. There is an opportunity to influence the standardisation of data capture and storage before a variety of new or competing methods, standards, and technologies become established. The adoption of existing data models, collection practices and the use of stable global identifiers as well as embedded metadata, quality and sharing capabilities into new and revised collection systems.
11. Improvements in environmental data use and re-use can be facilitated by a legislative framework and associated policies that mandate the data to be captured during relevant activities and, if available, to ensure that these data are provided to the relevant national facilities.
12. It is vital not to diminish the resources available to primary sources of data. An aggregator does not generate data and is dependent on the primary sources. The more those sources are able to engage in their core activities and collect data, the more that these data can be made available. Aggregators should at least assist custodians by providing useful data-sharing infrastructure so that they can direct their resources to focus on their core activities rather than resolving IT problems.

The unique nature of biodiversity data means that the role of aggregators such as the Atlas of Living Australia and the Terrestrial Ecological Research Network will be increasingly important. In addition to providing a central coordination role around discovery and access, they also fulfil a critical role regarding the archiving of a rich national record of environmental observations. Warehousing or aggregating such data also provides tight data integration and therefore supports the widest range of applications. Data integration also provides major computational advantages in regards to efficient data discovery, access and re-use. Major opportunities in understanding system dynamics and response also arise when exploring options for integrating data across domains, for example, coupling ecological observations with time series remote sensing. This is particularly important for Australian biodiversity data where national longitudinal time series data are lacking and surrogates, such as variables derived from time series satellite and other forms of imagery, are likely to play an increasingly important supporting role.

The study has traversed a suite of components necessary to establish a national biodiversity profiling capability and provided some high-level recommendations that have emerged from the study. Although of direct relevance to biodiversity profiling, the lessons from this study are also valuable for other environmental domains where understanding continental-scale environmental change is important.

8. Acknowledgments

A number of people provided direct support to this project and their contribution is greatly valued, including Matt Bolton (Department of Sustainability, Environment, Water, Population and Communities), Fiona Dickson (Department of Sustainability, Environment, Water, Population and Communities), Peter Doherty (ALA), John LaSalle (ALA), Warwick McDonald (Bureau of Meteorology) and Martin Wardrop (Department of Sustainability, Environment, Water, Population and Communities). We also acknowledge Trevor Booth and Andy Sheppard (CSIRO Ecosystem Sciences) and Michael Vardon (Australian Bureau of Statistics) for their comments on an earlier version of this report. We are grateful to the Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education for provision of data to support elements of this study, in addition to the National Collaborative Research Infrastructure Strategy through their ongoing support of the Atlas of Living Australia and related capabilities.

9. References

- Allnutt, T.F., Ferrier, S., Manion, G., Powell, G.V. N., Ricketts, T. H., Fisher, B.L., Harper, G.J., et al. 2008. A method for quantifying biodiversity loss and its application to a 50-year record of deforestation across Madagascar. *Conservation Letters*, 1, 173–181.
- Ashcroft, M.B., Gollan, J.R., Faith, D.P., Carter, G.A., Lassau, S.A., Ginn, S.G., Bulbert, M.W. and Cassis, G. 2010. Using Generalised Dissimilarity Models and many small samples to improve the efficiency of regional and landscape scale invertebrate sampling. *Ecological Informatics*, 5, 124–132.
- Australian Bureau of Agricultural and Resource Economics and Sciences. 2011. *Guidelines for land use mapping in Australia: principles, procedures and definitions*, fourth edition, Australian Bureau of Agricultural and Resource Economics and Sciences, Canberra.
- Belbin, L. 1987. The use of non-hierarchical allocation methods for clustering large sets of data. *Australian Computer Journal*, 19, 32–41.
- Belbin, L. 2011. The Atlas of Living Australia's Spatial Portal. In, M.B. Jones and C. Gries (eds.), *Proceedings of the Environmental Information Management Conference 2011 (EIM2011)*, 39–43. University of California, doi:10.5060/D2NC5Z4X, http://www.ala.org.au/wp-content/uploads/2011/10/EIM_ALA.pdf, Last Accessed, July 27, 2012.
- Berry, S.L. and Roderick, M.L. 2002a. Estimating mixtures of leaf functional types using continental-scale satellite and climatic data. *Global Ecology and Biogeography*, 11, 23–39.
- Berry, S.L. and Roderick, M.L. 2002b. CO₂ and land-use effects on Australian vegetation over the last two centuries. *Australian Journal of Botany*, 50, 511–531.
- Berry, S.L. and Roderick, M.L. 2006. Changing Australian vegetation from 1788 to 1988: effects of CO₂ and land-use change. *Australian Journal of Botany*, 54, 325–338.
- Booth, T.H., Williams K.J. and Belbin, L. 2012. Developing biodiverse plantings suitable for changing climatic conditions 2: Using the Atlas of Living Australia. *Ecological Management and Restoration*, 13, 274–281.
- Burley, H.M., Laffan, S.W. and Williams, K.J. 2012. Spatial non-stationarity and anisotropy of compositional turnover in eastern Australian Myrtaceae species. *International Journal of Geographical Information Science*, 26, 2065–2081.
- Executive Steering Committee for Australian Vegetation Information. 2003. *Australian Vegetation Attribute Manual: National Vegetation Information System, Version 6.0*. Canberra: Australian Government Department of the Environment and Heritage.
- Briggs, S. V. and Freudenberger, D. 2006. Assessment and monitoring of vegetation condition: Moving forward. *Ecological Management and Restoration*, 7, S74–S76.
- Bureau of Rural Sciences. 2010. User Guide and Caveats for the Land Use of Australia, Version 4, 2005–06. Bureau of Rural Sciences, Canberra.
- Bureau of Rural Sciences. 2006. User guide and caveats for the 1992/93, 1993/94, 1996/97, 1998/99, 2000/01 and 2001/02 land use of Australia, Version 3, Canberra, Department of Agriculture, Fisheries and Forestry, Australian Government.
- Bureau of Rural Sciences. 2009. User Guide and Caveats for the 2005/06 Land Use of Australia, Version 4, Canberra, Department of Agriculture, Fisheries and Forestry, Australian Government.
- Bureau of Rural Sciences. 2004. Land Use Mapping for the Murray-Darling Basin: 1993, 1996, 1998, 2000 maps. Australian Government Department of Agriculture, Fisheries and Forestry: Canberra.
- Bureau of Rural Sciences. 2006. Australia's Plantations 2006. Multi-Criteria Analysis Shell for Spatial Decision Support Datapack 1km. Canberra: Australian Government Department of Agriculture, Fisheries and Forestry.
- Cook, C.N. et al. 2010. Is what you see what you get? Visual vs. measured assessments of vegetation condition. *Journal of Applied Ecology*, 47, 650–661.
- Department of Climate Change and Energy Efficiency. 2012. Forest Extent and Change (v8). Area-Corrected, Aggregated Products. Canberra: Australian Government Department of Climate Change and Energy Efficiency (DCCEE).
- Department of the Environment and Water Resources. 2007. Australia – Present Major Vegetation Subgroups – NVIS Stage 1, Version 3.1 – Albers. Canberra: Australian Government Department of the Environment and Water Resources (DEWR).
- Drielsma, M. and Ferrier, S. 2006. Landscape scenario modelling of vegetation condition. *Ecological Management and Restoration*, 7, (Suppl. 1), S45–S52.
- Faith, D.P. and Ferrier, S. 2002. Linking beta diversity, environmental variation, and biodiversity assessment. *Science* 296, <http://www.sciencemag.org/cgi/eletters/295/5555/636#504>.

Ferrier, S. 2011. Extracting more value from biodiversity change observations through integrated modelling. *BioScience*, 61, 96–97.

Ferrier, S. 2012. Big-picture assessment of biodiversity change: scaling up monitoring without selling out on scientific rigour. Pp 63-70 in D. Lindenmayer and P. Gibbons, eds. *Biodiversity Monitoring in Australia*. CSIRO Publishing.

Ferrier, S., Manion, G., Elith, J. and Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, 13, 252–264.

Ferrier, S., Powell, G.V.N, Richardson, K.S., Manion, G., Overton, J.M., Allnutt, T.F., Cameron, S.E., Mantle, K., Burgess, N.D., Faith, D.P, Lamoreux, J.F., Kier, G., Hijmans, R.J., Funk, V.A., Cassis, G.A., Fisher, B.L., Flemons, P., Lees, D., Lovett, J.C. and van Rompaey, R.S.A.R. 2004. Mapping more of terrestrial biodiversity for global conservation assessment. *BioScience*, 54, 1101–1109.

Fitzpatrick, M.C., Sanders, N.J., Ferrier, S., Longino, J.T., Weiser, M.D. and Dunn, R. 2011. Forecasting the future of biodiversity: a test of single- and multi-species models for ants in North America. *Ecography*, 34, 836–847.

Flemons, P. and Belbin, L. 2010. Report on the environmental data library workshop, April 22-23 2010, online: http://www.ala.org.au/wp-content/uploads/2010/08/Lee-NEDL_Workshop-Report-Final.pdf, last accessed 30 March, 2013, Atlas of Living Australia.

Flemons, P., Raymond, B., Brenton, P. and Belbin L. 2010. ALA Spatial Analysis Tools Workshop Report, 3–4 December 2009, version 3.0. pp Page, online: http://www.ala.org.au/wp-content/uploads/2010/09/Lee_ALA-Spatial-Analysis-Tools-Workshop-Report-V3.pdf, last accessed March 30, 2013, Atlas of Living Australia.

Furby, S. 2002. *Land Cover Change: Specification for Remote Sensing Analysis*. Canberra: Australian Greenhouse Office.

GBIF. 2010. GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network, version 1.0. authored by Hill, A. W., Otegui, J., Ariño, A. H., and R. P. Guralnick. 2010. Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Accessible online at <http://www.gbif.org>. (<http://www.gbif.org/communications/news-and-events/showsingle/article/gbif-position-paper-enhancing-fitness-for-use-across-gbif>)

Leaper, R., Hill, N.A., Edgar, G.J., Ellis, N., Lawrence, E., Pitcher, C.R., Barrett, N.S. and Thomson, R. 2011. Predictions of beta diversity for reef macroalgae across southeastern Australia. *Ecosphere*, 2, art73.

Leathwick, J.R., Snelder, T., Chadderton, W.L., Elith, J., Julian, K. and Ferrier, S. 2011. Use of generalised dissimilarity modelling to improve the biological discrimination of river and stream classifications. *Freshwater Biology*, 56, 21–38.

Lymburner L., Tan P, Mueller N., Thackway R., Lewis A., Thankappan M., Randall L., Islam A. and Senarath U. 2011. *The National Dynamic Land Cover Dataset*, National Earth Observation Group, Geoscience Australia, Canberra.

Marsh, C.J., Lewis, O.T., Said, I. and Ewers, R.M. 2010. Community-level diversity modelling of birds and butterflies on Anjouan, Comoro Islands. *Biological Conservation*, 143, 1364–1374.

Miles, N. 2011. *Guide to data quality, version 1.2*. Atlas of Living Australia, Canberra. 17pp

Mutendeuzi, M. and Stafford-Bell, R. 2011. *Scientific information for making decisions about natural resource management – a report of the value, status and availability of key ABARES datasets*, ABARES technical report 11.2, Canberra, July.

National Reserve System. 2010. *Collaborative Australian Protected Areas Database – CAPAD 2008 (digital spatial data)*. Canberra: National Reserve System – Parks Australia, Department of Sustainability, Environment, Water, Population and Communities: Australian Government.

Overton, J.M., Barker, G.M. and Price, R. 2009. Estimating and conserving patterns of invertebrate diversity: a test case of New Zealand land snails. *Diversity and Distributions*, 15, 731–741.

Phillips, S.J. Dudík, M. and Schapire, R.E. 2004. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 655–662 http://www.cs.princeton.edu/~schapire/papers/maxent_icml.pdf Last Accessed, July 27, 2012.

Roderick, M.L., Noble, I.R. and Cridland, S.W. 1999. Estimating woody and herbaceous vegetation cover from time series satellite observations. *Global Ecology and Biogeography*, 8, 501–508.

Rosauer, D., Ferrier, S., Williams, K.J., Manion, G., Keogh, S. and Laffan, S.W. 2013. Phylogenetic Generalised Dissimilarity Modelling: a new approach to analysing and predicting spatial turnover in the phylogenetic composition of communities. *Ecography*, in press.

Tann, J., Kelly, L and Flemons, P. 2008. Atlas of Living Australia User Need's Analysis. <http://www.ala.org.au/wp-content/uploads/2010/07/ALAUUserNeedsAnalysisReportExtract.pdf> Last Accessed, July 27, 2012.

Thackway, R. and Lesslie, R. 2006. Reporting vegetation condition using the Vegetation Assets, States and Transitions (VAST) framework. *Ecological Management and Restoration*, 7 (Suppl. 1), S53–S62.

Thackway, R. and Lesslie, R. 2008. Describing and mapping human-induced vegetation change in the Australian landscape. *Environmental Management*, 42, 572–590.

Thomassen, H.A., Buermann, W., Milá, B, Graham, C.H., Cameron, S.E., Schneider, C.J., Pollinger, J.P., Saatchi, S., Wayne, R.K. and Smith T.B. 2010. Modeling environmentally associated morphological and genetic variation in a rainforest bird, and its application to conservation prioritization. *Evolutionary Applications*, 3, 1–16.

Wang, K., Franklin, S.E., Guo, X.L. and Cattet, M. 2010. Remote sensing of ecology, biodiversity and conservation: a review from the perspective of remote sensing specialists. *Sensors*, 10, 9647–9667

Williams, K.J., Belbin, L., Austin, M.P., Stein, J., and Ferrier, S. 2012. Which environmental variables should I use in my biodiversity model? *International Journal of Geographic Information Sciences*, 26, 2009–2047.

Williams, K.J., Ferrier, S., Rosauer, D., Yeates, D., Manion, G., Harwood, T., Stein, J., Faith, D.P., Laity, T., and Whalen, A. 2010a. *Harnessing Continent-Wide Biodiversity Datasets for Prioritising National Conservation Investment*. A report prepared for the Department of Sustainability, Environment, Water, Population and Communities, by CSIRO Ecosystem Sciences: Canberra. Available online at, <https://publications.csiro.au/rpr/pub?list=BRO&pid=csiro:EP102983>

Williams, K.J., Ferrier, S., Rosauer, D., Yeates, D., Manion, G., Harwood, T., Stein, J., Faith, D.P., Laity, T., and Whalen, A. 2010b. *Harnessing Continent-Wide Biodiversity Datasets for Prioritising National Conservation Investment: Appendices*. A report prepared for the Department of Sustainability, Environment, Water, Population and Communities, by CSIRO Ecosystem Sciences: Canberra. Available online at, <https://publications.csiro.au/rpr/>

[pub?list=BRO&pid=csiro:EP102983](https://publications.csiro.au/rpr/pub?list=BRO&pid=csiro:EP102983)

Williams, K.J., Reeson, A.F., Drielsma, M.J., and Love, J. 2012. Optimised whole-landscape ecological metrics for effective delivery of connectivity-focused conservation incentive payments. *Ecological Economics*, 81, 48–59.

Zerger, A., P. Gibbons, S. Jones, S. Doyle, J. Seddon, S. V. Briggs and D. Freudenberger. 2006. Spatially modelling native vegetation condition. *Ecological Management & Restoration*, 7 (s1), S37–S44.

Zerger, A, Gibbons, P., Seddon, J., Briggs, S, and Freudenberger, D. 2009. A method for predicting native vegetation condition at regional scales. *Landscape and Urban Planning*, 91, 65–77.

Notes

Notes

Notes

